Lecture Notes in Social Networks Vol.1

Nasrullah Memon
Reda Alhajj
Editors

From Sociology to Computing in Social Networks

Theory, Foundations and Applications

SpringerWienNewYork

SpringerWienNewYork

Lecture Notes in Social Networks (LNSN)

Series Editors

Nasrullah Memon
University of Southern Denmark
Odense, Denmark

Reda Alhajj
University of Calgary
Calgary, AB, Canada

For further volumes:
www.springer.com/series/8768

Nasrullah Memon · Reda Alhajj
Editors

From Sociology to Computing in Social Networks

Theory, Foundations and Applications

SpringerWienNewYork

Editors

Nasrullah Memon
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
5230 Odense, Denmark
memon@mmmi.sdu.dk

Reda Alhajj
Department of Computer Science
University of Calgary
Calgary, AB, Canada
alhajj@ucalgary.ca

SpringerWienNewYork is a part of Springer Science + Business Media
springer.at

Typesetting: Camera ready by the authors
Data Conversion: le-tex publishing services GmbH, Leipzig, Germany
Printing and Binding: Stürtz, Würzburg, Germany

Printed on acid-free paper
SPIN 80015037

With numerous (partly coloured) Figures

Library of Congress Control Number: 2010929763

ISSN 2190-5428
ISBN 978-3-7091-0293-0 SpringerWienNewYork

Contents

List of Contributors

Muhaimenul Adnan Department of Computer Science, University of Calgary, Calgary, Alberta, Canada;
email: adnanm@cpsc.ucalgary.ca

Reda Alhajj Department of Computer Science, University of Calgary, Calgary, Alberta, Canada;
Department of Computer Science, Global University, Beirut, Lebanon;
Department of Information Technology, Hellenic American University, Athens, Greece;
email: alhajj@ucalgary.ca

Sheeraz Ali Cursor Software Solutions, UAE;
email: sheeraz@cursorsoft.net

Ken Barker University of Calgary, Department of Computer Science, Calgary, Alberta, Canada;
email: kbarker@ucalgary.ca

Stephan Baumann German Research Center for AI (DFKI), Berlin, Germany;
email: stephan.baumann@dfki.de

Stephen Bird Clarkson University, Potsdam NY, USA;
email: sbird@clarkson.edu

Steffen Blaschke University of Hamburg, Germany;
email: steffen.blaschke@wiso.uni-hamburg.de

Neil Bozowsky Oculus Info Inc.;
email: nbozowsky@oculusinfo.com

Justus Bross Hasso-Plattner-Institute, Internet Technologies and Systems, Prof.-Dr.-Helmert-Strasse 2-3, 14482 Potsdam, Germany;
email: justus.bross@hpi.uni-potsdam.de

Jean-Hugues Chauchat ERIC Laboratoire-Université Lumière Lyon 2, Université de Lyon, France;
email: jean-hugues.chauchat@univ-lyon2.fr

Jiyang Chen Department of Computing Science, University of Alberta, Canada;
email: jiyang@cs.ualberta.ca

Marc Cheong Clayton School of Information Technology, Monash University, Victoria, 3800 Australia;
email: marc.cheong@infotech.monash.edu.au

Amir H. Chinaei University of Puerto Rico (Mayagüez campus), Department of Electrical and Computer Engineering, Mayaguez, PR, USA;
email: ahchinaei@ece.uprm.edu

Bruce Cronin University of Greenwich Business School, Park Row, London, UK;
email: b.cronin@greenwich.ac.uk

Elizabeth M. Daly IBM, Dublin Software Lab;
email: elizabeth_daly@ie.ibm.com

Kambiz Ghazinour University of Calgary, Department of Computer Science, Calgary, Alberta, Canada;
email: kghazino@ucalgary.ca

Randy Goebel Department of Computing Science, University of Alberta, Canada;
email: goebel@cs.ualberta.ca

Abdelghani Guerbas Department of Computer Science, University of Calgary, Calgary, Alberta, Canada

Andreas Harrer Katholische Universität Eichstätt-Ingolstadt, Eichstätt, Germnay;
email: andreas.harrer@ku-eichstaett.de

Carina Heppke University of Cambridge, Judge Business School Trumpington Street, Cambridge, UK;
email: ch471@cam.ac.uk

Panagiotis Karampelas Hellenic American University, Athens, Greece;
email: pkarampelas@hau.gr

Keivan Kianmehr Department of Computer Science, University of Calgary, Calgary, Alberta, Canada;
email: mkkian@ucalgary.ca

Iyakutti Kombiah CSIR Emeritus Scientist, School of Physics, Madurai Kamaraj University, Madurai, India;
email: iyakutti@gmail.com

Vincent Lee Clayton School of Information Technology, Monash University, Victoria, Australia;
email: vincent.lee@infotech.monash.edu.au

Cheng-Te Li National Taiwan University, Taipei, Taiwan;
email: d98944005@csie.ntu.edu.tw

Shou-De Lin National Taiwan University, Taipei, Taiwan;
email: sdlin@csie.ntu.edu.tw

Maryam Majedi University of Calgary, Department of Computer Science, Calgary, Alberta, Canada;
email: mmajedi@ucalgary.ca

Christoph Meinel Hasso-Plattner-Institute, Internet Technologies and Systems, Prof.-Dr.-Helmert-Strasse 2-3, 14482 Potsdam, Germany;
email: office-meinel@hpi-web.de

Nasrullah Memon Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark;
email: memon@mmmi.sdu.dk

Mohamad Nagi Department of Computing, School of Computing Informatics and Media, University of Bradford, Bradford, UK

Darko Obradović German Research Center for AI (DFKI) and University of Kaiserslautern, Germany;
email: darko.obradovic@dfki.uni-kl.de

Sampson Pun University of Calgary, Department of Computer Science, Calgary, Alberta, Canada;
email: szypun@ucalgary.ca

Pir Abdul Rasool Qureshi Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark;
email: parq@mmmi.sdu.dk

Balaji Rajendran Centre for Development of Advanced Computing, Bangalore, India;
email: balajirajendran@gmail.com

Dominique Raynaud PLC, Université de Grenoble / Université Paris Sorbonne, Paris, France;
email: dominique.raynaud@upmf-grenoble.fr

Keven Richly Hasso-Plattner-Institute, Internet Technologies and Systems, Prof.-Dr.-Helmert-Strasse 2-3, 14482 Potsdam, Germany;
email: keven.richly@student.hpi.uni-potsdam.de

Mick Ridley Department of Computing, School of Computing Informatics and Media, University of Bradford, Bradford, UK;
email: m.j.ridley@bradford.ac.uk

Jon Rokne Department of Computer Science, University of Calgary, Calgary, Alberta, Canada;
email: rokne@ucalgary.ca

Mike Savigny Oculus Info Inc.,
email: msavigny@oculusinfo.com

Patrick Schilf Hasso-Plattner-Institute, Internet Technologies and Systems, Prof.-Dr.-Helmert-Strasse 2-3, 14482 Potsdam, Germany;
email: patrick.schilf@student.hpi.uni-potsdam.de

David Schroh Oculus Info Inc.;
email: dschroh@oculusinfo.com

Asadullah Shaikh Universitat Oberta de Catalunya, Barcelona, Spain;
email: ashaikh@uoc.edu

Anna Stavrianou ERIC Laboratoire - Université Lumière Lyon 2, Université de Lyon, France;
email: anna.stavrianou@univ-lyon2.fr

Klaus Stein Laboratory for Semantic Information Technology, University of Bamberg, Germany;
email: klaus.stein@uni-bamberg.de

Julien Velcin ERIC Laboratoire-Université Lumière Lyon 2, Université de Lyon, France;
email: julien.velcin@univ-lyon2.fr

Uffe Kock Wiil Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark;
email: ukwiil@mmmi.sdu.dk

William Wright Oculus Info Inc.;
email: bwright@oculusinfo.com

Osmar R. Zaïane Department of Computing Science, University of Alberta, Canada;
email: zaiane@cs.ualberta.ca

Sam Zeini Universität Duisburg-Essen, Germany;
email: zeini@collide.info

Sabrina Ziebarth Universität Duisburg-Essen, Germany;
email: ziebarth@collide.info

Social Networks: A Powerful Model for Serving a Wide Range of Domains

Nasrullah Memon and Reda Alhajj

Abstract By nature, social networks exist and evolve; they are dynamic. However the development in information technology increased the interest in social networks and they are adapted to more applications and domains. To cope with the change, we took the initiative and started three major projects, conference, journal and lecture notes series that are intended to meet the expectations of the involved research groups whose members have diverse backgrounds. This book is the first volume in our new series entitled lecture notes in social networks. It includes papers that cover different topics ranging from fundamentals to applications of social networks.

1 General Overview

Naturally, since their existence on earth humans tend to come together and socialize leading to social communities that could dynamically change by having new members joining and some of the existing members leaving. It is very common for people to move from one country to another and even from one location to another within the same country. People may change political parties. Employees may change employer and even department within the same organization. Actually, we witness the establishment of social commu-

Nasrullah Memon
The Maersk Mc-Kinney Moller Institute,
University of Southern Denmark, Odense, Denmark.
e-mail: memon@mmmi.sdu.dk

Reda Alhajj
Dept of Computer Science, University of Calgary, Calgary, Alberta, Canada.
Dept of Computer Science, Global University, Beirut, Lebanon.
e-mail: alhajj@ucalgary.ca

nities as part of our daily life. This natural phenomenon has been realized by researchers in sociology early in the 20th century.

The new field of study required multidisciplinary knowledge covering in addition to psychology and sociology, statistics, linear algebra and graph theory, among others. For long time, the whole social network methodology was dependent on manual process and hence concentrated on small networks involving humans [1, 2, 3, 4, 6, 8, 9, 14, 17]. The main theme was to analyze human interactions in a certain environment in order to discover key persons, groups, etc.

The recent development in information technology and the wide spread of the world wide web have highly influenced and shaped the research on social networks [7, 12, 16]. People are joining online social networks and socialize on the web. There is a clear shift from real to virtual life.

Researchers are becoming more aware of the social network methodology as a model that could be utilized as a novel perspective to tackle research problems in different domains ranging from web mining and advertisement to the study of terrorist groups [2, 5, 6, 8, 10, 11, 13, 14, 15, 18]. The chapters of this book cover a wide range of domains as evidence on the applicability and importance of the social network methodology.

2 The Need for the Lecture Notes in Social Network Series

The mount of research outcome and the number of research groups working on social networks has rapidly increased over the past decade due the recognition of the social network methodology as an effective model to develop unique solutions for some vital problems. As a result, it is becoming difficult for existing forums and venues to meet the expectations of the diverse population of researchers working in this multidisciplinary area.

Realizing the need though somehow late, we took the initiative and started three major projects that could partially cover the visible gap due to the shift into automated tools and techniques to handle and analyze social networks. Our target has always been to produce multidisciplinary venues that satisfy the expectations of researchers from all disciplines which have already realized the importance of social networks. It is also necessary to keep in mind the necessity to motivate for new application areas and domains. We are aware of the need to work hard in order to succeed in achieving the mission. Starting a project could be feasible but sustainability should be the target. In 2007, we started putting together a major international conference in social network. We wanted it be a real multidisciplinary venue and to avoid duplicating an existing venue; it was not easy to establish the team and find sponsors. The first International Conference on Advances on Social Network Analysis and Mining (ASONAM) was hosted by the Hellenic American University in Athens,

Greece in July 2009. The acceptance rate was below 30% and the conference was well attended. Participants from North America, Europe, Asia and the Middle East enjoyed the scientific and social programs of ASONAM 2009. The second version of the conference, ASONAM 2010 has been scheduled to take place in Odense Denmark in August 2010 and ASONAM 2011 will be held in Kaohsiung Taiwan. Thanks to Hellenic American University, The University of Calgary, University of Southern Denmark, Global University, Mehran University of Engineering and Technology, Springer, IEEE and ACM for their generous support without which it would have been hard to realize ASONAM as a leading conference. Following the successful organization of ASONAM 2009, we negotiated two other major initiatives with Springer, to start a new journal and book series. Special thanks go to Stephen Soehnlen who has highly motivated us and always pushed us hard to turn the dream into reality. The Social Networks Analysis and Mining Journal web set was activated in December 2009. We have the pleasure to put together an international committee of associate editors consisting of elite researchers in their disciplines. The first issue of the journal is schedule to appear in Fall 2010.

Starting the new series "Lecture Notes in Social Networks" (LNSN) is the last project that is officially activated by publishing this edited book as the first volume in the series.

3 Organization of the Volume

This book provides a look at some of the latest research results in a variety of specialized topics that are central to the current research trends in social networks; it is the first volume of the Lecture Notes in Social Networks series.

This volume is composed of 21 contributions authored by some of the prominent researchers. The focus of this volume is multidisciplinary and hence it is entitled: "From Sociology to Computing in Social Networks: Theory, Foundations and Applications". The contributions span a wide variety of technical areas within this research field. The chapters of this volume have been organized into five categories, namely mining based social network methods, dynamics in social network models, discovering structure in social networks, social media, and software applications. A brief overview of the contributions in each of those categories is provided in the sequel.

Section 1: Mining-based Social Network Methods

The first section of this volume begins with a contribution by Nagi et al. In this chapter, authors presented an integrated framework that combines the power of social network methodology and web mining techniques in or-

der to produce a comprehensive and robust approach capable of effectively optimizing websites for better navigation. The authors integrated the three web mining techniques, namely web structure, usage and content mining and also utilized the web log in order to construct the social network of the pages constituting the website under investigation. For constructing the social network, the authors demonstrated the power of association rules mining by considering user sessions as transactions and the pages themselves as the items. The results obtained from the social network based methodology are consistent with the results reported by the web structure based techniques, namely PageRank and HITS. The authors validated the method using the data set which shows that this is a simple but viable approach to solve the given problem. The next contribution in this section is from Li and Lin and in this chapter authors presented a method for egocentric information abstraction for heterogeneous social networks. The method can be applied to create a node-based search engine for social networks as well as realizing social network visualization.

The first section continues with a contribution from Stavrianou, Velcin, and Chauchat who described a theoretical work that consists in defining formally a Post-Reply Opinion Graph. The main novelty of the contribution is integrating into the model structure information of the discussion and the opinion content of the exchanged forum postings, information that is lost when authors represent a forum by a social network model. Authors define measures that give information regarding the opinion flow and the general attitude of users and towards users throughout the whole forum. The application of the proposed model to real forums shows the additional information that can be extracted and the interest in combining the social network and the PROG models.

The third chapter by Rajendran and Kombiah constructs a contextual semantic structure by observing the actions of the users involved in Web-based Knowledge Gathering (WKG) task, in order to gain an understanding of their task and requirement. The authors also build a knowledge warehouse in the form of a master Semantic Link Network (SLN) that accommodates and assimilates all the contextual semantic structures. This master SLN, which is a socio-contextual network, is then mined to provide contextual inputs to the current users through their agents. In the chapter, the authors validated their approach through experiments and analyzed the benefits to the users in terms of resource explorations and the time saved. In the final contribution of the first section, Adnan, Alhajj, and Rokne proposed a social network modeling technique that takes as input the data to be analyzed for constructing a social network and maps it into a new space by mining frequent closed patterns. Using the produced frequent closed patterns, the authors create a useful set of features to represent an entity that describes the connection of the entity to data in a reasonable way. The results presented in this chapter also verify conjecture for the synthetic and real datasets.

Section 2: Dynamics in Social Network Models

A contribution by Harrer, Zeini, and Ziebarth begins the second section of the book. It demonstrates a visualization method to augment sociograms representing network communities with information about the temporal aspects and dynamics of the community. The authors created a three-dimensional representation that can be manipulated interactively to boards, and bibliography sources for an automatic transformation into social network data formats. In the next contribution of the section, Qureshi, Memon, and Wiil present a model and system architecture for an early warning system to detect terrorist threats. The chapter discusses the shortcomings of state-of the-art systems and outlines the functional requirements that should be met by an ideal system working in the counterterrorism domain. The concept of generation of early warnings to predict terrorist threats is presented in the chapter. The presented model relies on data collection from open data sources, information retrieval, information extraction for preparing structured workable data sets from available unstructured data, and finally detailed investigation. The conducted investigation includes social network analysis, investigative data mining, and heuristic rules for the study of complex covert networks for terrorist threat indication. The presented model and system architecture can be used as a core framework for an early warning system.

The second section continues with a contribution from Cronin examines the roll-out of an electronic knowledge base in a medium-sized professional services firm over a six year period. The efficiency of such implementation is a key business problem in IT systems of this type. The analysis provides some evidence of mathematical complexity in the periodicity. Some implications of complex patterns in social network data for research and management are discussed. The study provides a case study demonstrating the utility of the broad methodological approach. In the final contribution of the section, Daly presents reputation ranking which measures the overall popularity of a bookmark, taking into account the timely relevance of the document. The ranking mechanism presented is relatively simple, and involves little computational overhead by using a basic reward/decay model. The technique couples the reputation of users with the reputation of the document being bookmarked in an attempt to capture the "wisdom of the crowds". As a result, reputable users, e.g., trend-setters have a greater influence on the reputation of the bookmarked documents than users with a low reputation value.

Section 3: Discovering Structures in Social Networks

The third section starts with a contribution by Chen, Zaïane, and Goebel that review problems of existing methods for constructing local communities, and propose a new metric to evaluate local community structure when

the global information of the network is unavailable. Based on the metric, the authors develop a two-phase algorithm to identify the local community of a set of given starting nodes. The method does not require arbitrary initial parameters, and it can detect whether a local community exists or not for a particular node. The authors have tested the algorithm on real world networks and compared its performance with previous approaches. Experimental results of the contribution confirm the accuracy and the effectiveness of the metric and algorithm.

The third section continues with a contribution from Raynaud that scrutinizes network forcing of diffusion process. The departure of empirical data from the logistic function is explained by social network discreteness, heterogeneity and anisotropy. New indices are proposed in the contribution and results are illustrated by empirical data from an original study of knowledge diffusion in the medieval academic network. In the following contribution, Stein, and Blaschke discuss precise measures of collaborative intensity with respect to not only the width but also the depth of collaboration. Based on empirical data of four social networks, the authors compare a widely-used approximation with their own measures of collaborative intensity. The authors also find that the quality of the approximation varies with the type of social networks. In the final contribution of the section, Bird investigates the relationship between the centrality of individual actors in a social network structure and their policy learning performance. This analysis demonstrates that an informational environment in which one links to their less embedded network alters has a strong impact within collaborative-based learning processes.

Section 4: Social Media

The fourth section starts with a contribution by Obradović, and Baumann in which they developed and applied an efficient variation of core-analysis to blog data-sets of different languages. The analyses revealed a general tendency towards more-than-average core-centralization, as well as a number of interesting phenomenon, that were not all publicly known by now. Mining, analyzing, modeling and presenting vast pool of knowledge in one central framework to extract, exploit and represent meaningful knowledge for the common blog user forms the basis of the next contribution by Bros et al. The result of the corresponding long-term research initiative presented in the contribution is BLOG INTELLIGENCE. It is an integrated blog analysis framework with the objective to leverage content- and context-related structures and dynamics residing in the blogosphere and to make these findings available in an appropriate format to anyone interested. The authors refer to these structures and dynamics as social physics of the blogosphere. The next contribution in this section is from Cheong and Lee and presents

a framework for discovery of the emergent properties of users of the Twitter micro-blogging platform. The novelty of the contribution is the use of machine-learning methods to deduce user demographic information and online usage patterns and habits not readily apparent from the raw messages posted on Twitter. This is different from existing social network analysis performed on de facto social networks such as Facebook, in the sense that authors use publicly available metadata from Twitter messages to explore the inherent characteristics about different segments of the Twitter community, in a simple yet effective manner. The framework presented in the chapter is coupled with the self-organizing map visualization method, and tested on a corpus of messages which deal with issues of socio political and economic impact, to gain insight into the properties of human interaction via Twitter as a medium for computer-mediated self-expression. A contribution by Cheong and Lee concludes the section with a review on the state of the art in literature regarding Twitter as a micro-blogging service. This review has hopefully given the reader an insight into potential research on Twitter, specifically bridging the gap between the user/message domains on Twitter as the majority of existing research deals with the messages (tweets) on Twitter with little emphasis given on the underlying users behind them.

Section 5: Software Applications

A contribution by Heppke begins the fifth section of the book. This conceptual article takes a social capital perspective in order to explain the benefits and challenges of social software inside the firewall of organizations. Corporate social software is considered to hold great benefits for the management and the efficient use of knowledge within organizations which is regarded to become an increasingly important capability for companies in changing and challenging business environments in which adaptation, change and innovation are required to stay ahead. However, the extent to which the benefits of corporate social software are realized by organizations depends on the way that social technologies are actually used inside the firewall. While external social technologies such as Facebook and Twitter have quickly established themselves in the daily usage patterns of a large majority of people, the usage of similar technologies within the firewall of organizations is characterized by distinct differences which are discussed in this chapter.

The fifth section continues with a contribution from Ghazinour et al in which they describe an extension onto database catalogues and Structured Query Language (SQL) for supporting privacy in Internet applications, such as in social networks, e-health, e-government, etc. The idea is to introduce new predicates to SQL commands to capture common privacy requirements, such as purpose, visibility, generalization, and retention for both mandatory and discretionary access control policies. Furthermore, the extension is sim-

ple, modular, and backward compatible so it is applicable to new designs or legacy systems. The model is supported with semantics defined operationally in a relational model. In particular, relevant details affecting the underlying catalogues and their supporting algorithms are discussed.

The third contribution in this section is from Schroh et al.; the authors introduced the nCompass framework and integration platform. The framework describes key nCompass core services, and provides results on functional synergies achieved through technology service integration with nCompass. In the final contribution of the fifth section, Shaikh et al. present present a secure web-based architectural design by using the standards of Service Oriented Architecture (SOA) for distributed web application that maintains the interoperability and data integration through certain secure channels. The authors have created CRUD (Create, Read, Update, Delete) operations that has an implication on their own created web services and we propose a secure architecture that is implemented on CRUD operations. The contribution provides an extensive description of the prevention of replay attacks and a detailed explanation for applying security measures.

Acknowledgements

The editors would like to gratefully acknowledge the efforts of all those who have helped create this volume. Firstly, it would never be possible for a volume of book such as this one to provide such a broad and extensive look at the latest research in the field of social network without the efforts of all those expert researchers and practitioners who have authored and contributed papers. Their contributions made this volume possible.

In addition, we would like to thank the reviewers for their time and effort in the preparation of their thoughtful reviews. Their support was crucial for ensuring the quality of this volume and for attracting wide readership.

Moreover, we would like to thank Hellenic American University, University of Southern Denmark, and University of Calgary for their support, and encouragement. We are also grateful for the pleasant cooperation with Stephen Soehnlen and his team from Springer and their professional support in publishing this volume. Lastly, but certainly not least, the editors would like to acknowledge the considerable effort and support provided by Asadullah Shaikh with the layout, typesetting, and formatting of the book.

References

1. Albert, R. and Barabási, A.L. Statistical mechanics of complex networks. *Reviews of Modern Physics,* 74(1):47-97, 2002.

2. Barabási, A.L. and Albert, R. Emergence of scaling in random networks. *Science,* 286(5439):509-512, 1999.
3. Baumes J., Goldberg M., Magdon-Ismail M. and Wallace W., "Discovering hidden groups in communication networks," *Proceedings of NSF/NIJ Symposium on Intelligence and Security Informatics,* 2004.
4. Backstrom L., Huttenlocher D., Kleinberg J. and Lan X., "Group formation in large social networks: Membership, growth, and evolution," *Proceedings of ACM KDD,* 2006.
5. Croft D. P., James R., Thomas P., Hathaway C., Mawdsley D., Laland K. and Krause J., "Social structure and co-operative interactions in a wild population of guppies (poecilia reticulata)," *Behavioural Ecology and Sociobiology,* Vol.59, No.5, pp.644-650, 2006.
6. Domingos, P. and Richardson, M. Mining the network value of customers. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* San Francisco, CA: ACM Press, 2001.
7. Flake, G.W., Lawrence, S. et al. Efficient identification of web communities. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Boston, MA: ACM Press, 2000.
8. Janssen, M.A. and Jager, W. Simulating market dynamics: Interactions between consumer psychology and social networks. *Artificial Life,* 9:343-356, 2003.
9. Jensen D. and Neville J., "Data mining in social networks," *Proceedings of the Symposium on Dynamic Social Network Modeling and Analysis,* 2002.
10. Kianmehr K. and Alhajj R., "Calling Communities Analysis and Identification Using Machine Learning Techniques," *Expert Systems with Applications,* Vol.36 , No.3, pp.6218-6226, 2009.
11. Klerks P., "The Network Paradigm Applied to Criminal Organisations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands," *CONNECTIONS,* Vol.24, No.3, pp.53-65, 2001.
12. Lawrence, S. and Giles, C.L. Accessibility of information on the web. *Nature,* 400: 107-109, 1999.
13. Memon N. and Larsen H. L., "Structural Analysis and Mathematical Methods for Destabilizing Terrorist Networks," *Proceedings of the International Conference on Advanced Data Mining Applications,* Springer-Verlag Lecture Notes in Artificial Intelligence (LNAI 4093), pp.1037-1048, 2006.
14. Menczer, F. Evolution of document networks. *Proceedings of the National Academy of Science of the United States of America,* 101:5261-5265, 2004.
15. Newman, M.E.J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Science of the United States of America,* 98:404-409, 2001.
16. Pennock, D.M., Flake, G.W. et al. Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Science of the United States of America,* 99(8):5207-5211, 2002.
17. Powell, W.W., White, D.R. et al. Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences.*American Journal of Sociology,* 110(4):1132-1205, 2005.
18. Xu, J.J. and Chen, H. CrimeNet Explorer: A framework for criminal network knowledge discovery. *ACM Transactions on Information Systems,* 23(2):201-226, 2005.

Part I
Mining-based Social Network Methods

Employing Social Network Construction and Analysis in Web Structure Optimization

Mohamad Nagi, Abdelghani Guerbas, Keivan Kianmehr, Panagiotis Karampelas, Mick Ridley, Reda Alhajj, and Jon Rokne

Abstract The world wide web is growing continuously and rapidly; it is quickly facilitating the migration of tasks of the daily life into web-based. This trend shows time will come when everyone is forced to use the web for daily activities. Naive users are the major concern of such a shift; so, it is necessary to have the web ready to serve them. We argue that this requires well optimized websites for users to quickly locate the information they are looking for. This, on the other hand, becomes more and more important due to the widespread reliance on the many services available on the Internet nowadays. It is true that search engines can facilitate the task of finding the information one is looking for. However, search engines will never replace but do complement the optimization of a website's internal structure based on previously recorded user behavior. In this chapter, we will present a novel

Mohamad Nagi
Department of Computing, School of Computing Informatics & Media, University of Bradford, Bradford, UK

Abdelghani Guerbas
Department of Computer Science, University of Calgary, Calgary, Alberta, Canada

Keivan Kianmehr
Department of Computer Science, University of Calgary, Calgary, Alberta, Canada

Panagiotis Karampelas
Department of Information Technology, Hellenic American University, Athens, Greece

Mick Ridley
Department of Computing, School of Computing Informatics & Media, University of Bradford, Bradford, UK

Reda Alhajj
Department of Computer Science, University of Calgary, Calgary, Alberta, Canada
Department of Computer Science, Global University, Beirut, Lebanon
Department of Information Technology, Hellenic American University, Athens, Greece

Jon Rokne
Department of Computer Science, University of Calgary, Calgary, Alberta, Canada

approach for identifying problematic structures in websites. This method consists of two phases. The first phase compares user behavior, derived via web log mining techniques, to a combined analysis of the website's link structure obtained by applying three methods leading to more robust framework and hence strong and consistent outcome: (1) constructing and analyzing a social network of the pages constituting the website by considering both the structure and the usage information; (2) applying the Weighted PageRank algorithm; and (3) applying the Hypertext Induced Topic Selection (HITS) method. In the second phase, we use the term frequency-inverse document frequency (TFIDF) measure to investigate further the correlation between the page that contains the link and the linked to pages in order to further support the findings of the first phase of our approach. We will then show how to use these intermediate results in order to point out problematic website structures to the website owner.

1 Introduction

The rapid development in the technology has motivated for a shift towards electronic documentation and communication. Almost everything is turning into web-based service and more sophisticated websites are being developed in a very competitive environment. Actually, website complexity is constantly increasing, making it more difficult for users to quickly locate the information they are looking for. This, on the other hand, becomes more and more important due to the widespread reliance on the many services available on the Internet nowadays. Analyzing the user's behavior along with the internal website structure and content will provide insight on how to optimize the website's structure in order to improve its usability. This is possible and feasible by a combined approach that integrates web mining techniques with social network analysis. We are mainly motivated by the successful application of the social network methodology in different domain, e.g., [14, 18, 28].

Web Mining is the use of data mining methods to identify patterns and relationships amongst web resources. It is basically classified into: web content, web usage and web structure mining; these are used to solve the website structure optimization problem. Web structure mining involves the crawling and analysis of web page content to identify all links existing within the page, which will then be used to generate a directed graph representing the structure of the site being mined. Each node within the produced graph signifies an individual page and each edge is a link between two pages. On the other hand, web usage mining requires the parsing of web server logs to identify individual user behavior. Specifically, the sites visited, total visits and total time spent looking at the page, also known as "think time", are considered. These values are parsed from original server logs, or could be taken from preprocessed logs as well. Web content mining, on the other hand, helps in

determining the anticipated correlation between the context around the link and the page to which the link points. Finally, social network analysis is intended to combine the web usage information into the process in order to find a network structure alternative to the one produced by web structure mining. Analyzing both networks could lead to more strong and consistent results. This is true because the combined approach investigates the website structure from two perspectives, the actual structure and the structure induced from web usage.

In this chapter, we describe a novel approach for web structure optimization. The proposed approach involves two phases. The first phase uses web structure and web usage mining combined with social network construction and analysis techniques in order to find an initial recommendation for restructuring the analyzed website. The second phase relies on web content mining to determine a more appropriate linking between pages by considering the context around each link. The combined result from both phases leads to a higher confidence in the final recommendation; we use the term frequency inverse document frequency measure (TFIDF) [27, 16, 26], which is the most common measure used in analyzing documents for information retrieval.

For the first phase, we applied two trends. The first trend investigates how to use both Hypertext Induced Topic Selection (HITS) method [21] and the Weighted PageRank algorithm [3, 5, 29] for web-structure mining; this lead to structure based analysis of the hyperlink structure of a website. We derive another structure which is realized as a social network constructed by considering the web log content. Actors in the social network are the pages constituting the website and links are derived by applying association rules mining technique on the web log content. Two pages are linked together if they, respectively, appear in the antecedent and consequent of an association rule. The weight of the link increases to reflect the number of supporting association rules.

An association rule reflects the correlation between sets of items (hereafter called itemsets); items in our model are pages. Association rules mining is a two step process, first frequent itemsets are found and then association rules are generated. An itemset is frequent if it is supported by a certain percentage of the transactions larger than a prespecified minimum threshold value (mostly specified by an expert). In our settings, transactions are user sessions that include a listing of the pages visited by the user. Once all frequent itemsets are identified, each frequent itemset is used to derive all possible rules such that the antecedent and consequent of the rule are disjoint and their union leads to the itemset. A rule is acceptable if it has high confidence, i.e., the support of all its items divided by the support of the items in the antecedent is larger than a predefined minimum confidence values (mostly specified by a domain expert).

We further demonstrate how to use web log mining to obtain data on the site user's specific navigational behavior. We then describe a scheme, how to interpret and compare these intermediate results to measure the web-

site's efficiency in terms of usability. Based on this, it shall be outlined how to make recommendations to website owners in order to assist them in improving their site's usability. It is obvious that such an approach is needed to make websites more attractive to end users who navigate websites looking for particular information. Having the information buried deep within the website discourages users from continuing their link navigation process. They may stop visiting such websites and move to the competitors; a situation not appreciated by website owners. Therefore, the result from the approach described in this chapter is intended to make websites more attractive to most users by considering users' behavior and website structure. The reported test results demonstrate the effectiveness and applicability of the proposed approach. Finally, it is worth emphasizing that search engines will never be considered alternative to the website optimization process described in this chapter. Rather search engines could complement this process once a user is not willing to navigate through the pages of different websites. The user in such a case may prefer to use a search engine as a first step to land on a particular page and from there navigate to other pages.

To sum up, the contributions of the work described in this chapter could be summarized as follows.

- Benefiting from web structure, web log and web content information in order to analyze the structure of a website.
- Utilizing the social network methodology to derive an alternative structure by analyzing the web log.
- Combining all the results into an integrated robust approach that leads to more consistent and stronger recommendations for altering the current structure of the analyzed website.

The rest of this chapter is organized as follows. Section 2 is related work. Section 3 describes the proposed approach; we first present how web structure mining is utilized in the process of website optimization and describe the participation of web usage mining to the process; then we discuss how web content mining helps in increasing the confidence in the recommendation; and finally, we discuss how the overall recommendation is conveyed to the user of the analyzed website. Section 4 reports test results that demonstrate the applicability and effectiveness of the proposed integrated approach. Section 5 is summary and conclusions.

2 Related Work

As described in the literature, numerous approaches have been taken to analyze a website's structure and correlate these results with usability, e.g., [8, 9, 11, 12, 13, 20, 23]. For instance, the work described in [22] devised a spatial frequent itemset data mining algorithm to efficiently extract

navigational structure from the hyperlink structure of a website. Navigational structure is defined as a set of links commonly shared by most of the pages in a website. The approach is based on a general purpose frequent itemset data mining algorithm, namely ECLAT [7]. ECLAT is used to mine only the hyperlinks inside a window with adaptive size; it slides along the diagonal of the website's adjacency matrix. They compared the results of their algorithm with results from a user-based usability evaluation. The evaluation method gave certain tasks to a user (like for example finding a specific piece of information on a website) and recorded the time needed to accomplish a task and failure ratios. The researchers found a correlation between the size of the navigational structure set and the overall usability of a website, specifically the more navigational structure a website has, the more usable it is as a general rule of thumb. In our work described in this chapter, we use frequent itemset data mining algorithm, namely FP-growth [2] to serve a purpose different than that described in [22]; the latter mines hyperlinks inside a window; it is a kind of web structure mining for a partition of the website (specific number of pages that are closely linked together). On the other hand, our study is comprehensive and covers the whole website; we mainly derive a social network between the pages based on their appearance in the association rules. Further, ECLAT could be effective for small scale mining tasks; but it is not scalable because it inverts the data and does operations directly on the columns. FP-growth [2] is more scalable; our research group have developed an effective disk-based version that could well overcome the memory limitations problem.

The work described in [25] analyzes the web log using data mining techniques to extract rules and predict which pages users will be going to visit based on their prior behavior. It is then shown how to use this information to improve the website structure. By using data mining techniques, this approach is also somehow related to our approach described in this chapter, although the details of the method vary greatly, due to their use of frequent itemset data mining algorithms to predict future usage. The main difference between our approach and the method described in [25] is that they do not consider the time spent on a page by a visitor in order to measure the importance of that particular page. Their approach applies frequent itemset mining that discovers navigation preferences of the visitors based on the most frequent visited pages and the frequent navigational visiting patterns. However, we believe that in a particular frequent navigational pattern there might exist some pages which form an intermediate step on the way to the desirable page that a user is actually interested in. Therefore, the time spent on a page by a visitor has to be considered as an important measure to quantify the significancy of a page in a website structure. Further, we do not limit our investigation to individual paths navigated by users, we rather use the outcome from the mining process to decide on the links between the pages in the constructed social network; the more rules favor a specific link the higher weight it gets and hence the most central it turns in the analysis of the social

network. Therefore, our results reflect a global picture of the website usage and navigational patterns.

The work described in [17] proposed two hyperlink analysis-based algorithms to find relevant pages for a given Web page. The work is different in nature from our work; however it applies web mining techniques. The first algorithm extends the citation analysis to web page hyperlink analysis. The citation analysis was first developed to classify core sets of articles, authors, or journals to different fields of study. In the context of Web mining, the hyperlinks are considered as citations among the pages. The second algorithm makes use of linear algebra theories to extract more precise relationships among the Web pages in order to discover relevant pages. By using linear algebra, they integrate the topological relationships among the pages into the process in order to identify deeper relations among pages for finding the relevant pages. The work described in [15] describes an expanded neighborhood of pages with the target to include more potentially relevant pages.

In [30], the authors outline a method of preparing web logs for mining specific data on a per session basis. This way, an individual's browsing behavior can be recorded using the time and page data gathered. Preparations to the log file such as stripping entries left by robots are also discussed.

In the approach described in [29], the standard PageRank algorithm was modified by distributing rank amongst related pages with respect to their weighted importance, rather than treating all pages equally. This results in a more accurate representation of the importance of all pages within a website. We used the Weighted PageRank formula outlined in [29] to complement the web structure mining portion of our approach, with the hope of returning more accurate results than the standard PageRank algorithm. The result obtained from the weighted PageRank is first validated by applying HITS [21] to check the consistency of the results, and then confirmed or complemented by the outcome from the employed social network methodology. HITS is another document ranking algorithm developed by Kleinberg; it also ranks documents based on link information.

3 The Proposed Website Analysis Approach

Hyperlinks encode what developers anticipate as the desired connection between pages of a website as to be navigated by end users. Specifically, the existence of a link from page p to page q on the Web represents a concrete indication that the two pages p and q contain some related information; such pages are mostly visited together, which is mostly not true given the diversity of web users who may have different interests. Consequently, it is more appropriate to revisit, evaluate, and adjust a website design periodically by considering users' behavior as reflected in the web log and the website structure which is recognized as a directed graph. The other directed graph con-

sidered and analyzed in this chapter is a social network of the pages with the links derived from the web log data.

In general, a collection V of hyperlinked pages can be viewed as a directed graph $G = (V, E)$, where the nodes correspond to the pages, and a directed edge $(p, q) \in E$ indicates the presence of a link from page p to page q. Each web page presented in the graph has an out-degree and in-degree; each of the two degrees may be any positive integer greater than or equal to zero. The out-degree of a page is the number of links in the page; such links lead to other mostly related pages. The in-degree of a page p is the number of pages which include a hyperlink to page p. So, our target is to check the links and validate them by applying a systematic approach that incorporates certain criteria to evaluate the overall structure of a website.

In order to achieve our goal of recommending changes to the link structure of a website, we have identified two main subproblems which must be initially solved. The first subproblem involves determining which pages are important, as implied by the structure of the website; two sources are used for this information: the actual hyperlink structure and the social network structure. The actual hyperlinks structure is analyzed using two methods, namely HITS and weighted PageRank. The other structure is derived as a social network by using the web log. The current website structure is said to satisfy the visitors if the social network based structure is very similar to the actual hyperlinks structure. In other words, the social network reflects the trends in users' navigational patterns and the target is to adjust the current website structure to best fit users' expectations by matching the actual structure of the website to the deduced social network structure. The outcome from this first subproblem complements the investigation of the second subproblem where main target is to conclude which pages the users of the studied website consider to be important, based on the information amassed from the web log.

From the above description of the two subproblems, it is obvious that the web log serves two purposes for solving the two subproblems. Once we have solved these two subproblems, we now have methods in place which give us two different rankings of the same web pages. Our final task is implementing a scheme to compare the results of the first two perspectives and make meaningful recommendations. In the subsections which follow, we will discuss the algorithms we will use for unravelling each of these tasks and the reasoning behind these algorithms. The outcome from this first phase is supported by a second phase that emphasizes web content mining to study the correlation around each link in a page and the pages directly and indirectly pointed to by the current page.

In the rest of this section, we describe the different components of the first phase of the proposed approach, namely web structure and web log mining; we then analyze the combined result from the first phase. Finally, we concentrate on the second phase which completes the overall analysis process.

3.1 Web Structure Mining

Web structure mining involves crawling through a series of related web pages (for example all pages inside a user defined subdomain), extracting meaningful data that identifies the page, and use that data to give the page a rank based on given criteria. The crawling process proceeds as follows. First, a set or the root of web pages is provided. Second, an application called a crawler will traverse the set of pages or start at the given root page and extract the needed information from all the visited pages. The information we are interested in for the work described in this chapter are the hyperlinks contained within the visited pages. This is the basic information for producing the graph summarizing the website structure.

It is worth noting that the same information can be extracted using regular expressions. However, there are challenges using regular expressions because they assume that the code used within the web pages follows all standards. Simple errors, such as some HTML tags not being closed or improperly formatted and non-HTML code, such as CSS or javascript, can throw off the parsing of the page and this may lead to inaccurate results.

3.1.1 Weighted PageRank Based Ranking

Once the hyperlinks within the visited web page have been extracted, the crawler will recursively continue crawling web pages whose links were found in the already visited web pages; all replicates are removed, since crawling the same page twice is unnecessary. The process will end up removing any duplicate hyperlinks within the visited page as well as any links that have already been processed or are already in the queue awaiting processing. After having crawled the complete website, or a user defined part of it, depending on what the user specifies, the standard page rank value $PR(p_i)$ of each page p_i can be computed as

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \tag{1}$$

where N is the total number of pages that have been crawled, $M(p_i)$ is the set of pages that link to p_i, $L(p_j)$ is the number of outgoing links on page p_j, and d is a damping factor, usually chosen around 0.85. The damping factor can be interpreted as the probability that a user follows the links on a page. It has been included due to the following observation: *Sometimes a user does not follow the links on a page p_k and just chooses to see a random page p_l by entering its address directly*; this could be seen as a good example of using a search engine to land at a particular page. Such a case should be considered when computing the page rank of p_l. Thus, in the above formula, $\frac{1-d}{N}$ can be seen as the influence of a random jump to page p_i based on the page rank

$PR(p_i)$. As a result, the above formula is comprehensive enough to cover all possible cases from navigating through pages of the website to landing directly at a particular page.

Xing and Ghorbani [29] proposed an improved version of standard page rank, namely the weighted page rank algorithm (WPR). Their basic goal is to consider the fact that the page rank of a popular page should have a higher weight than the page rank of an unpopular page. The WPR value is computed as:

$$WPR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} PR(p_j) W^{in}_{(p_j,p_i)} W^{out}_{(p_j,p_i)} \tag{2}$$

Here, $W^{in}_{(p_j,p_i)}$ and $W^{out}_{(p_j,p_i)}$ are the weights of the link between documents p_j and p_i. They can be computed as:

$$W^{in}_{(p_j,p_i)} = \frac{I_{p_i}}{\sum_{p \in R(p_j)} I_p} \tag{3}$$

$$W^{out}_{(p_j,p_i)} = \frac{O_{p_i}}{\sum_{p \in R(p_j)} O_p} \tag{4}$$

where I_x is the number if links pointing to page x, O_x is the number of outgoing links in page x and $R(x)$ is the set of pages that are linked to/from page x. Each element of the sum of page ranks is multiplied by it's respective weight. The result leads to the more realistic fact that more important pages are given a higher page rank, unlike the original page rank algorithm that divides the rank of a page evenly amongst documents it links to. At the end, a more consistent, realistic and accurate result is achieved.

After thorough analysis of the possible techniques [19], we decided to use WPR instead of the standard page rank, since it has proven to yield slightly better results in experiments as reported by Xing and Ghorbani [29]. Using this concept also fills the gap between the structure based analysis and the social network based analysis because the latter methods derives a weight per link by considering association rules. In other words, the outcome from the two methods will be comparable without any of the two methods taking advantage over the other; both are deriving links between pages of the website and both assign weight to each derived link. However, structure derived by the social network methodology is more practical because it reflects the actual links by analyzing the web log. On the other hand, the web structure mining method derives all the available links regardless of whether they have been used on not; every hyperlink leads to an edge in the derived structure and this edge may not have a corresponding edge in structure due to the social network. We decided to assign zero weight to all such links unless they are highly favored by weight from the weighted PageRank algorithm. By assigning zero weight, we drop from further consideration every link not favored by the web log and hence the integration of the results becomes smoother.

The output of the weighted PageRank based processing stage is a list where each entry contains a web page and its weight computed by the weighted PageRank, denoted $[p_i, WPR(p_i)]$. Entries in the list are sorted in descending order by the weight values, WPR. However, the list is revised by reordering the entries based on the outcome from applying another web pages ranking algorithm, namely HITS [21] as well as the social network methodology.

3.1.2 HITS Based Ranking

HITS is another algorithm for ranking web pages based on two values for each page, namely `authority` and `hub`. Authority refers to pages that provide an important, trustworthy information on a given topic; and hub are pages that contain links to authorities. These two values are defined in terms of one another in a mutually reinforcing relationship: a better hub points to many good authorities, and a better authority is pointed to by many good hubs. In other words, authority of a page p is computed as the sum of the scaled hub values that point to page p; and hub of page p is computed as the sum of the scaled authority values of the pages points to by page p. Relevance of the linked pages is also considered in some implementations. In-degree of a page measures its authoritativeness; and the out-degree measures the hubness. Hubs and authorities together form a bipartite graph.

Classifying pages into authority and hubs directly influences the result from the weighted PageRank algorithm by favoring more pages classified as authorities.

3.1.3 Social Network Based Ranking

The social network is built by analyzing the web log data. Actors in the social networks are the pages. We construct the adjacency matrix by mining association rules from the transactional database obtained after preprocessing the web log data; each transaction is a set of pages accessed together in one session. First, all entries in the adjacency matrix are set to zero. Then, FP-growth is applied on the derived transactional data and association rules are derived. Each rule is reflected in the adjacency matrix by incrementing every entry (i, j) such that pages i and j exist in the antecedent and consequent of the rule, respectively. Finally, entries in the adjacency matrix are normalized by dividing each value by the overall average of the values that exist in the matrix. The outcome is used as input to Pajek (a social networks construction and analysis tool) [4]. The social network is analyzed to rank the pages by considering their in-degrees, out-degrees, and betweenness centrality. Pages with high between centrality are considered as important to link pages from different communities.

Thee measures obtained from the HITS algorithm and the social networks methodology are used in combination with the outcome from the weighted PageRank algorithm to produce a final ranking for the pages; the process is mostly important to break ties and to validate the ranking of pages by recomputing the rank for each page as the average rank from the three approaches. This turns the page ranking process into a more stable task with high confidence in the final ranked list.

3.2 Ranking Pages Based on Web Log Mining

Seeking user opinion is always a target but mostly hard to achieve thoroughly. However, we discovered that the web log could be the best available source that may be used to pull out users' opinions without hurting any party and even without any precautions that might lead to bias. Users normally access websites with certain target in mind; they navigate between pages at their own choice; they decide how long to spend on each page and they may decide to leave and come back later. All this valuable trace is recorded and summarized in the web log. Therefore, we can analyze the web log and extract users' opinions which could be another metric for ranking the web pages. We base the latter ranking on two parameters: (1) frequency (number of visits), and (2) time (total time spent by all users at a web page). These are important factors which could be computed by analyzing the web log.

3.2.1 Preprocessing:

The web log is in general a summary of all the activities done by the users visiting the website. It is necessary to clean the web log before it is used; this step is referred to as preprocessing. The target of the preprocessing step is to "clean" the web log in order to minimize interference from robots; this step will make the web log ready to generate the actual output of this stage. Actually this is the preprocessing step for the whole system, but we delayed its discussion until this point because it is more related to web log processing which is the core of the material discussed in this section. So, the input to the social network methodology is the clean web log data produced after this preprocessing step.

One approach we consider useful for this preprocessing step has been proposed in [30], where it is proposed to discard those sessions that match the following access patterns that are likely to be robots characteristics:

- Visiting around midnight, during light traffic load periods in order to avoid time latency.

- Using `HEAD` instead of `GET` as the access method to verify the validity of a hyperlink; the HEAD" method performs faster in this case as it does not retrieve the web document itself.
- Doing breadth search rather than depth search; robots do not navigate down to low-level pages because they do not need to access detailed and specific topics
- Ignoring graphical content; robots are not interested in images and graphic files because their goal is to retrieve information and possibly to create and update their databases.

Based on the cleaned log-file, we then identify sessions, which in turn are used to compute the total number of visits v_i and the total time spent by users, t_i for each page p_i. It shall be noted that we must ensure that the number of user sessions we extract from the web log are of sufficient size to give us a realistic ranking of popular pages.

3.2.2 Computing Log Rank Values

The first parameter to consider in the log ranking process is the number of times a particular page was visited by the group of users registered in the log. The fact that a user has visited a page may lead to the situation that the page is considered important. While this is true in many cases, there is also the possibility that the page was just an intermediate step or "hop" on the way to the page which the user is actually interested in. To illustrate this, consider three pages A, B and C, a very high number of visits from page $A \rightarrow B \rightarrow C$, and a relatively low total time spent at page B would imply that page B is used primarily (or even exclusively) as a "hopping" point. In this case, a viable recommendation may be to change the link structure of the website so that the user is able to navigate directly from $A \rightarrow C$, without having to make a stop in-between at page B. Therefore, we need a way to give lesser weight to the visit counter and a higher weight to the total time in our ranking scheme.

Assuming that v_i is the number of visitors for a page i and t_i is the total time spent by all visitors at page i, the *log rank* value l_i shall be defined as:

$$l_i = 0.4v_i + 0.6t_i \tag{5}$$

Here it is important to note that the number of visitors to a given page p is computed by summing all occurrences of page p in the log file. In other words, in case a user visited page p say three times then v_i of page p is incremented by 3, not by 1. This way, we are able to include in the log rank a measure of how important page p is to every user. Such a measure will be lost when the number of visits by each user is normalized down to 1 for visited pages and to 0 for unvisited pages. Actually, this ranking is directly reflected on to the social network of the web pages because the number of times a web page is

visited is taken as its frequency each time a sequence of pages constituting a session are added to the FP-tree.

Taking a weighted sum of the number of visits and time spent at the page will result in a value that represents the importance of a particular page relative to the other pages. That is, pages that are frequently visited and accessed for long periods of time will have a larger log rank than pages with an insignificant number of visits and think time. Rather than giving time and visits equal importance as discussed above, the difference is quantified experimentally through a constant, in this case being a 60/40 split, respectively. Depending on the content of the web site being analyzed, these constants can be changed to account for the specific audience or purpose of the site. For example, a website with pages normally filled with large amounts of visual/textual information will have on average longer think times than pages sparsely filled with content. A balance between having fewer pages with large amounts of content, versus many pages with little content must be identified and quantified in the log rank function; this will lead to a more accurate function that better reflects the particular case being analyzed. Finally, the output of this stage, is a list of pages sorted by their log rank value l_i.

3.3 Analyzing the Outcome from the First Phase

This is the last stage of processing; it yields results directly for the users in the form of recommendations on how to change their websites. The required input for this stage includes:

- A sorted list of pages with their page rank values $p_i \in R^+$ (the list is sorted by the page rank values). This is the combined rank produced by the three approaches weighted PageRank, HITS and social network.
- A sorted list of websites with their ranking values $l_i \in R^+$ from the weblog mining (the list is sorted by the ranking values).

These two criteria are utilized by the three steps of the analysis as described in the remainder of this section.

3.3.1 Preprocessing

First, we need to preprocess the page rank and log ranking values. This step consists of normalizing the two values to a common index of integers. This is done by taking the list of page ranks and sorting them in descending order. We then chose to assign a nondecreasing unique index to each page rank value, i.e., the largest page rank receives index zero and the smallest page rank receiving the highest index value. Recall that one of the main targets of

developing the integrated approach that employs the weighted PageRank, the social network methodology, and the HITS algorithm in the process has been to get a more robust framework that produces more consistent results. In case of a tie in the rank, then the page that has higher rank according to the social network methodology is given preference. We decided on giving higher priority to the social network result because it combines both the structure and web log features of the website. Finally, the same normalization process is repeated for the log ranking values.

This step transforms and maps the log and page rank values from two different distributions into a simple linear distribution, and hence facilitates appropriate comparison. We will validate this step during the evaluation of our method (see Section 6), where the log and page rank values have significantly different distributions, which would yield different analysis results. Finally, the output of this step consists of two lists, the page rank and log rank indexes along with their respective pages. At this step, the lists are still independent of each other.

3.3.2 Guideline to Recommendations

After the preprocessing step, we can now compute the following value d_i for each page:

$$d_i = index(l_i) - index(p_i) \tag{6}$$

Equation (6) computes the difference in rank between the two rank values. Ideally, we will find $d_i = 0$, because there should be little deviation in the ranking of the page rank and log rank values. This point has been indirectly raised above when we discussed the analysis of the social network of web pages. Here, it is worth highlighting that d_i will be zero when the page from gets same rank from weighted PageRank, HITS and the social network because the social network combines both the structure and web log information; it is actually another way of quantifying the value of d_i. Finally, the pages will be sorted in ascending order according to their d_i values. At the top and the bottom of the list, we can distinguish the following two cases, respectively:

Case 1 Pages that have got a high page rank index and a low log rank index (d_i very low).

Case 2 Pages that have got a low page rank index and a high log rank index (d_i very high).

In case 1, the software performing the analysis should recommend the user to put the page into a place where it is harder to reach, in favor of pages that might require to be reachable more easily. This includes but is not limited to:

- Removing links to such a page, especially links that appear in pages with high page rank.
- Linking to the page from places with low page rank value instead.

In case 2, the developed system should recommend modifying the link structure in a fashion that makes the popular page easier to reach. This means, for example, adding links to such popular pages, especially on suitable pages with high page rank, or may be direct jumping onto such popular pages from pages far towards the head along the link leading to the popular pages. In other words, shorten the path needed to be traversed in order to reach pages highlighted in case 2.

The intuition for a very high or very low deviation generally being undesirable, is the following: One could interpret a high page rank value as a page being easily reachable from other (important) pages, whereas a low page rank value thus could be interpreted as an indicator, that the page is hard to reach. On the other hand, a high log rank value testifies that a page is popular, whereas low log rank values indicate unpopular pages. Therefore, in the first case with a very low d_i, the site is easy to reach, but mostly few people actually want to see it; and in the second case with a very high d_i, the page is very hard to reach for visitors, but comparably many people want the information on it and have to spend time looking for it. Thus, it is natural, that according to this scheme, an ideally positioned page has a value $d_i \approx 0$.

3.4 Ranking Pages by Employing Web Content Mining

For this phase of the process, we decided on using the term frequency-inverse document frequency (TFIDF) measure, which has been successfully applied and produced promising results in information retrieval and text mining. Is is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. In other words if a term/word appears frequently in a document but also appears frequently in the corpus/collection as a whole, it will get a lower score. An example of this would be "the", "and", "it". However, depending on the source material, other words may be very common to the source matter.

The web content mining problem may be classified under text mining because we want to find the correlation between the term in the link and other documents directly or indirectly pointed to by the page. As TFIDF is used to evaluate how important a word is to a document in a collection or corpus, we could smoothly use the measure by considering the corpus in our case as the set of documents pointed to directly or indirectly by following a particular link within the given page. For the term constituting the link, its frequency is measured within the linked to page(s). Frequency of a term is measured

as the number of times the term appears in a document; it is normalized to avoid bias towards longer documents, which might have more occurrences of the given term as compared to shorter documents. Term frequency, denoted TF, of a term t_i in document p_j is measured by:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{7}$$

where $n_{i,j}$ is the number of occurrences of term t_i in document p_j, and the denominator is the size of document p_j, i.e., the number of occurrences of all terms in document p_j. The inverse document frequency, denoted IDF, measures the importance of term t_i. It is computed as the logarithm of the quotient obtained by dividing the number of all documents by the number of documents containing the term.

$$IDF_i = log(\frac{|D|}{|\{p_j : t_i \in p_j\}|}) \tag{8}$$

where $|D|$ is the total number of documents in the corpus, $|\{p_j : t_i \in p_j\}|$ is the number of documents in which term t_i appears (*i.e.*, $n_{i,j} \neq 0$).

Combining both values computed above, we get $TFIDF$ as:

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i$$

A high term frequency in the given document leads to a high weight value of $TFIDF$.

We applied this process to find the $TFIDF$ value for each term that appears in a hyperlink inside a page. The returned result guides the optimization process. It is a step to confirm or tune up the outcome from the first phase. It is considered an important step because the first phase considers only the structure and usage of the web pages. On the other hand, by computing the $TFIDF$ values for the hyperlinks, we dive deeper inside the web pages, analyze their content and comment on the importance of the link to the web page it points to as compared to the other web pages indirectly reached from the current page by following the hyperlinks.

3.5 The Relinking Process

Although we have a systematic approach that involves two phases to analyze a website seeking how its links could be optimized, our approach is semi-automated in the sense that user (expert) validation of the outcome is needed before the optimization is physically applied to the website. In particular, the proposed approach will highlight the restructuring decisions to be taken for a better optimized website. These decisions are to be investigated by domain experts and guide them for better restructuring the website. So, the aforementioned "relinking" process has to be carried out manually by the

website owner, who is considered the domain expert since he/she has to consider the content structure of the page; thus the scope of the proposed process to suggest relinking in terms of "Link page A to B" or "Remove the link on page A". The outlined process also represents a support in determining possibly misplaced pages and in deciding where to add or remove links (page rank values can be helpful here).

To assist him/her in the process, after this stage, the website owner should be presented with the following information:

- The sorted list of pages (called *UNLINK* list), with $d_i < 0$ and $d_i^* > \epsilon_1$, where ϵ_1 is a user defined threshold.
- The sorted list of pages (called *LINK-TO* list), with $d_i > 0$ and $d_i^* > \epsilon_2$, where ϵ_2 is another user defined threshold.
- For each web page in the above lists, provide the set of pages that link to it (incoming links) and the set of pages that are being linked to from it (outgoing links).
- The page rank and log rank values for each web page that has been analyzed, including but not limited to those in the *UNLINK* and the *LINK-TO* lists.
- The importance of each link by considering the $TFIDF$ values computed for the terms that appear in the hyperlinks within the pages. This is presented as pairs of hyperlink and related pages, where the pages are sorted by their $TFIDF$ values.

This information should be sufficient to detect and resolve design issues in a website's structure that affect usability. The ranking approach is supportive in that it helps the owner to focus on the important issues. To guide the process of relinking or altering the structure, page rank and log rank values are provided. The $TFIDF$ values combined with the page rank and log rank values will give better insight to the web site owner.

4 Evaluation of the Proposed Approach

We tested our approach on a medium sized website ($\approx$ 831 pages) obtained from [24], which provides reference for HiFi devices. Its structure is mostly wide rather than deep, as for example when it lists the manufacturers of documented devices. Since this website has been provided for experiments with data mining techniques, it already came with a log file that had been parsed into sessions. This saved us the time and effort required to derive the sessions from the raw web log.

We analyzed the sessions and produced the transactions which are lists of pages accessed per session. These transactions were used as the input to FP-growth which produced the association rules; the minimum support and minimum confidence thresholds were set to 20% and 75%, respectively. The

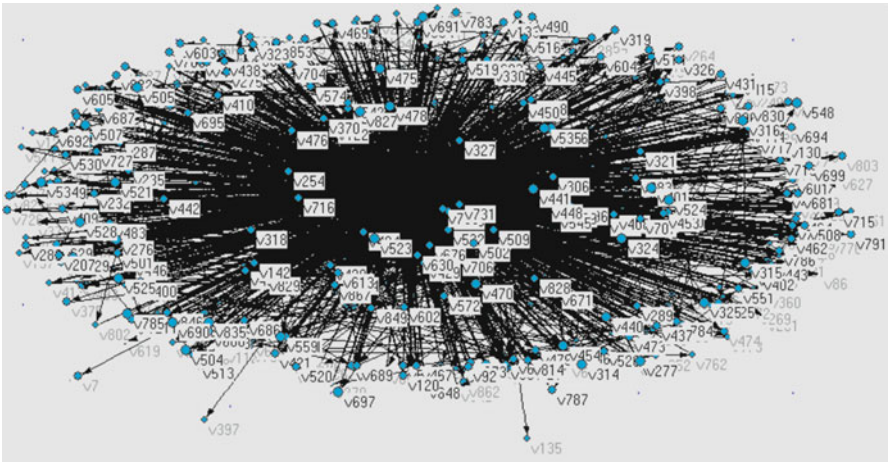


Fig. 1: The complete social network

association rules were then reflected into the adjacency matrix where entry (i, j) measures the number of rules that include the two pages i and j in the antecedent and consequent, respectively. The derived social network is shown in Figure 1. The social network of the top ranked pages is shown in Figure 2.

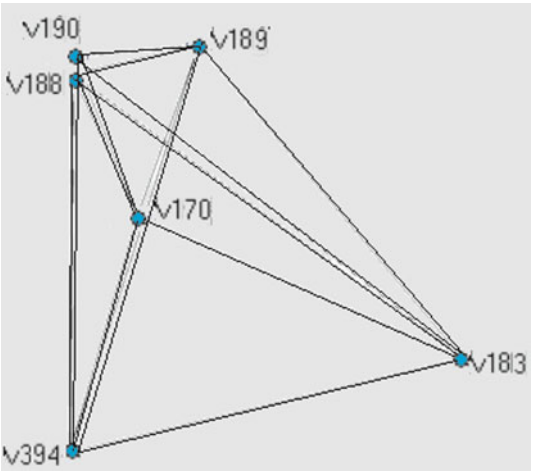


Fig. 2: The most central pages in gthe social network

The employed social network analysis technique has been used to rank the web site pages according to their importance; this could be used to validate the results outputted by PageRank and HITS. A centrality analysis has

been performed on the social network made of the website's pages and links between these pages. After comparing the results produced by the website structure based techniques (PageRank and HITS) with the results produced by the social network analysis based method we have found that both agree almost totally on ranking the most important pages. Therefore, the most important pages discovered by both approaches are almost the same. The top ranked pages are listed in Figure 3. For example: /manufacturers/index.html -394-, /guide/index.html -183- and /email.html -170- have been ranked exactly in the same place by both approaches and they are in both top lists. The other most important pages have close ranking but not exactly the same. The exception is the following page /analogue-heaven/index.html. It has been ranked as the sixth most important page by the website structure based techniques but it has not been considered as a central point according to the social network centrality measure. After closer analysis of this particular page, the explanation of this phenomenon could be articulated as follows: this specific page has two incoming edges from two of the very powerful pages. Therefore, the website structure based techniques have considered this page as very important because this is how the website structure based techniques work. On the other hand, this page was not considered as a central page by the social network analysis based method because it has very few incoming and outgoing edges.

Fig. 3: The top six pages as ranked by the first phase)

The combined approach of the first phase analysis on the site yielded interesting distribution of deviation values d_i. These results from the first phase of the proposed approach have been supported by consistent results from the second phase of the proposed approach. In other words, the computed d_i values and $TFIDF$ values are consistent for this particular website.

The reported values of d_i demonstrate that we have a relatively low number of pages with a deviation far from the ideal value. The majority of the pages fall within a small margin of ± 200, which is still acceptable. Some pages like for example /dr-660/index.html (has lowest d_i value) showed a large discrepancy between user popularity and reachability, since it was linked to from one of the central pages, but hardly received any hits. Other pages like /manufacturers/korg/s-3/index.html (has second highest d_i value) appear to

have been very popular for the site users, but are relatively hard to reach since they are hidden “deep" in the website’s structure. A viable change in this case would be to provide a link to it on the pages at or close to the website’s document root (for example in a “Favorites” or “Recommendations” section), since this is where the users start browsing. Further investigation of the highest and lowest values, showed the same tendency and thus revealed locations where relinking seemed necessary after manual investigation from our side.

To sum up, the integrated robust framework analyzed in this section is powerful enough to produce consistent and legitimate recommendations. It is a rich framework because it integrates three main pieces of information, namely web structure, web content and web log. The integration of the social network methodology into the framework is a major gain as reflected by the consistent results produced. The influence of the social network produced is high because it integrates both the structural and usage information together. In other words, it smoothly bridges the outcome from web structure and web log mining. As a result, the proposed framework has succeeded in identifying problematic locations in the website’s structure.

5 Summary and Conclusion

In this chapter, we presented an integrated framework that combines the power of social network methodology and web mining techniques in order to produce a comprehensive and robust approach capable of effectively optimizing websites for better navigation. We did not restrict the process to only web structure mining. We rather integrated the three web mining techniques, namely web structure, usage and content mining. We also utilized the web log in order to construct the social network of the pages constituting the website under investigation. For constructing the social network, we demonstrated the power of association rules mining by considering user sessions as transactions and the pages themselves as the items. The results is another perspective to view the web structure. Fortunately and surprisingly, the results obtained from the social network based methodology are consistent with the results reported by the web structure based techniques, namely PageRank and HITS. Finally, we use $TFIDF$ in order to analyze the correlation between the term in each hyperlink and the pages it points to directly or indirectly. Our approach then showed how to combine these values in order to measure a website’s usability. We successfully validated our method using the data set provided under [24], which shows that this is a simple but viable approach to solve the given problem. In our opinion, a similar method should be used as part of a larger set of tools, when it comes to usability optimization of websites. We will investigate how it would be possible to minimize the user involvement in the process by trying to automate the discovery of the

parameters and may be the final analysis of the results. An alternative immediate solution could be realized by presenting the results to domain experts in fuzzy terms as they are more understandable by humans. These issues are all needed to be investigated for their applicability and feasibility.

References

1. S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. *Proceedings of the International Conference on World Wide Web*, pp.280-290, 2003.
2. M. Adnan and R. Alhajj, "DRFP-Tree: Disk-Resident Frequent Pattern Tree," *Applied Intelligence,* Vol.30, No.2, pp.84-97, 2009.
3. A. Altman and M. Tennenholtz. Ranking systems: the pagerank axioms. *Proceedings of ACM Conference International on Electronic commerce*, pp.1-8, 2005.
4. V. Batagelj, A. Mrvar : *Pajek — Program for Large Network Analysis*. Home page: http://vlado.fmf.uni-lj.si/pub/networks/pajek/
5. M. Bianchini, M. Gori, and F. Scarselli. Inside pagerank. *ACM Transactions on Internet Technology*, 5(1):92–128, 2005.
6. P. Boldi, M. Santini, and S. Vigna. Pagerank as a function of the damping factor. *Proceedings of the International Conference on World Wide Web*, pp.557-566, 2005.
7. C. Borgelt. Efficient implementations of apriori and eclat, *Proceedings of the Workshop of Frequent Item Set Mining Implementations,* Melbourne, FL, 2003.
8. A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Transactions on Internet Technology*, 5(1):231–297, 2005.
9. J. T. Bradley, D. V. de Jager, W. J. Knottenbelt, and A. Trifunovic. Hypergraph partitioning for faster parallel pagerank computation. *Proceedings of Formal Techniques for Computer Systems and Business Processes, European Performance Engineering Workshop*, pp.155–171, 2005.
10. S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource compilation by analyzing hyperlink structure and associated text. *Proceedings of the International Conference on World Wide Web*, 1998.
11. Y.-Y. Chen, Q. Gan, and T. Suel. I/o-efficient techniques for computing pagerank. *Proceedings of ACM International Conference on Information and knowledge management*, pp.549–557, 2002.
12. P.-A. Chirita, J. Diederich, and W. Nejdl. Mailrank: using ranking for spam detection. *Proceedings of ACM International Conference on Information and knowledge management*, pp.373–380, 2005.
13. J. Cho, S. Roy, and R. E. Adams. Page quality: in search of an unbiased web ranking. *Proceedings of ACM SIGMOD*, pp.551–562, 2005.
14. L. da F. Costa, F. A. Rodrigues, G. Travieso and P. R. Villas Boas. Characterization of complex networks: a survey of measurements *Advanced Physics*, Vol. 56, pp 167–242, 2007.
15. J. Dean and M. Henzinger. Finding related pages in the world wide web. *Proceedings of the International Conference on World Wide Web*, 1999.
16. G. Guo, et al., An kNN Model-Based Approach and Its Application in Text Categorization. *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*, pp.559–570, 2004.
17. J. Hou and Y. Zhang. Effectively finding relevant web pages from linkage information. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):940–951, 2003.

18. W. H. Hsu, A. King, M. S. Paradesi, T. Pydimarri, and T. Weninger. Collaborative and structural recommendation of friends using weblog-based social network analysis. *In AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs (CAAW)*, volume SS-06-03, pages 55–60, Menlo Park, CA, 2006.
19. J. Jeffrey, P. Karski, B. Lohrmann, K. Kianmehr and R. Alhajj, "Optimizing Web Structures Using Web Mining Techniques," In Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning, Springer-Verlag LNCS, Brimingham, UK, 2007.
20. X.-M. Jiang, G.-R. Xue, W.-G. Song, H.-J. Zeng, Z. Chen, and W.-Y. Ma. Exploiting pagerank at different block level. *Proceedings of the International Conference on Web Information Systems Engineering*, pp.241–252, 2004.
21. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pap.668–677, 1998.
22. C.H. Li and C.K. Chui. Web structure mining for usability analysis. Proceedings of *IEEE/WIC/ACM International Conference on Web Intelligence*, pp.309–312, 2005.
23. P. Massa and C. Hayes. Page-rerank: Using trusted links to re-rank authority. *Proc. of IEEE/WIC/ACM International Conference on Web Intelligence*, pp.614–617, 2005.
24. U. of Washington Artificial Intelligence Research. Music machines website. http://www.cs.washington.edu/ai/adaptive-data/.
25. I. V. Renáta Iváncsy. Frequent pattern mining in web log data. *Journal of Applied Sciences at Budapest Tech*, 3(1):77–90, 2006.
26. P. Soucy and G. W. Mineau, Beyond TFIDF Weighting for Text Categorization in the Vector Space Model. *Proceedings of the International Joint Conference on Artificial Intelligence*, pp.1130–1135, 2005.
27. R. Steinberger, B. Pouliquen and J. Hagman, Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*, pp.415–424, 2002.
28. X. Wan, E. Milios, N. Kalyaniwalla, and J. Janssen, "Link-based event detection in email communication networks," in *SAC '09: Proceedings of the 2009 ACM symposium on Applied Computing*. New York, NY, USA: ACM, 2009, pp. 1506–1510.
29. W. Xing and A. A. Ghorbani. Weighted pagerank algorithm. *Proceedings of Annual Conference on Communication Networks and Services Research*, pp.305–314, 2004.
30. J. X. Yu, Y. Ou, C. Zhang, and S. Zhang. Identifying interesting customers through web log classification. *IEEE Intelligent Systems*, 20(3):55–59, 2005.

Mining Heterogeneous Social Networks for Egocentric Information Abstraction

Cheng-Te Li and Shou-De Lin

Abstract Social network is a powerful data structure that allows the depiction of relationship information between entities. However, real-world social networks are sometimes too complex for human to pursue further analysis. In this work, an unsupervised mechanism is proposed for egocentric information abstraction in heterogeneous social networks. To achieve this goal, we propose a vector space representation for heterogeneous social networks to identify combination of relations as features and compute statistical dependencies as feature values. These features, either linear or cyclic, intend to capture the semantic information in the surrounding environment of the ego. Then we design three abstraction measures to distill representative and important information to construct the abstracted graphs for visual presentation. The evaluations conducted on a real world movie dataset and an artificial crime dataset demonstrate that the abstractions can indeed retain significant information and facilitate more accurate and efficient human analysis.

1 Introduction

"Information abstraction" generally refers to the summarization and reorganization of the overwhelmed, raw information to a humanly-understandable representation while still retaining the important and meaningful messages. Figure 1 shows that overwhelming and complex information can usually hinder further manual analysis. In this work, we exploit the idea of information

Cheng-Te Li,
National Taiwan University,
e-mail: d98944005@csie.ntu.edu.tw

Shou-De Lin,
National Taiwan University,
e-mail: sdlin@csie.ntu.edu.tw

abstraction in heterogeneous social networks. Further, given the fact that a real-world social network can contain thousands or even millions of individuals and relations, and therefore users might not be interested in the network as a whole, rather they are particularly interested in the information of certain instances. Therefore, we propose the *egocentric* abstraction problem attempting to summarize the information of a given node. Borrowing from social network literatures [13], the node of interests can be referred as the *ego*. The ego node and its directly or indirectly connected neighbors compose a so-called *egocentric network*. The egocentric analysis highlights the micro view of the network. In other words, the information to be retained or discarded depends on the ego that users focus on. Thus, as will be shown in the evaluation, an egocentric abstraction can assist human in answering questions such as "which individual might be suspicious" or "*what is special about the specified movie star*" more efficiently.

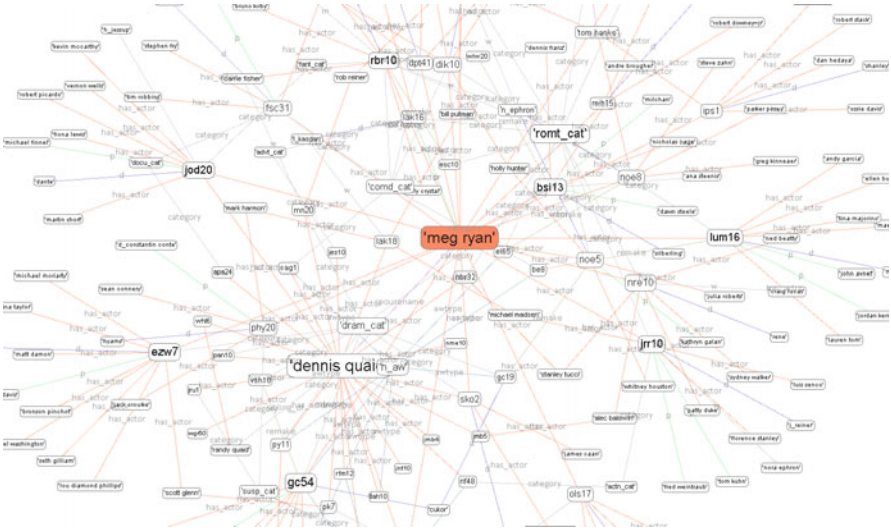


Fig. 1: The 2-neighborhood graph of "Meg Ryan."

One important characteristic of this study is that we pay special attention to the heterogeneous social networks (HSN) [13]. A heterogeneous social network contains a set of typed nodes (e.g. nodes can be movies, actors, or directors in the movie domain) and typed edges as relations (e.g. friends, family, and directs). Our goal is to perform the egocentric information abstraction in an HSN.

Although there are already various successful proposals on social network analysis (SNA), most assume there is only single type of nodes and single type of relations in a network. This kind of social network is defined as homogeneous social networks [13]. For example, both the Web and the citation

graph (i.e., nodes are authors and edges represent co-authorship) can also be regarded as a homogeneous social network because there is only one type of node (i.e., webpage or paper) and relation (i.e., hyperlink or citation link). However, in the real-world different types of objects can be connected through different kinds of relationships, therefore it is natural to define different types of entities and relations in a social network. In this sense, a more universal data structure, called heterogeneous social network, has been proposed to describe the complex relationships (i.e., a set of typed edges) among entities. For example, a heterogeneous movie network shown in Figure 2 takes movies (M), directors (D), writers (W), and actors (A) as nodes, and their corresponding relationships as tuples such as $< D_1, direct, M_1 >, < M_1, hasactor, A_1 >, < A_1, produce, M_1 >$ *and* $< M_3, originatefrom, M_4 >$, where the capital letter in the tuple stands for the type of source node, and the second element stands for the type of relations. Note that in general there could have multiple relationships in between two entities, like the relations of "*has actor*" and "*produce*" between A_1 and M_1.

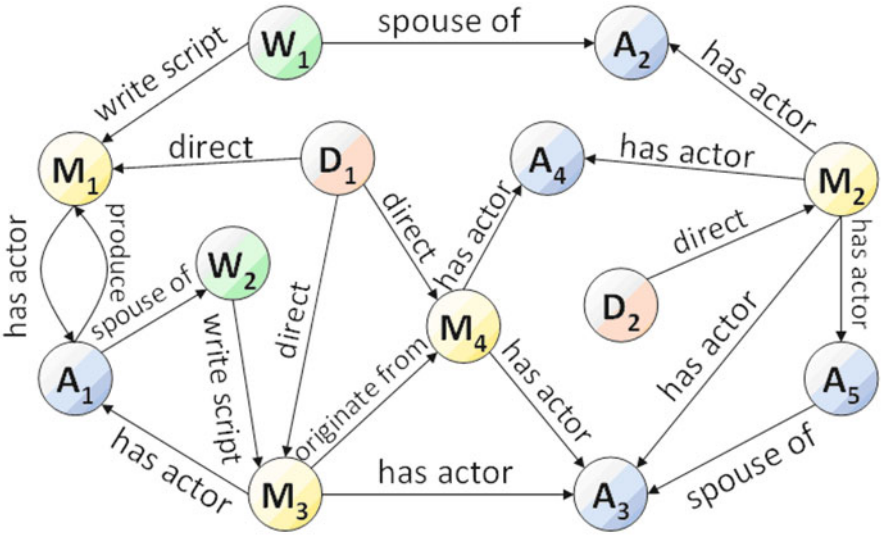


Fig. 2: A heterogeneous social network for movie domain. The capital letter of each node stands for its type: M(movie), D(director), A(actor), and W(writer). Besides, there are five relation types, including *"write script", "has actor", "spouse of", "direct", "produce",* and *"originate from"* in this example.

The concept of information abstraction has not yet been formally defined in heterogeneous social networks. Though the essences of several works are related to abstraction in some sense, they all suffer from a main deficiency for ignoring high-order relationship information. For example, centralities [4]

and PageRank [2] aim at finding important nodes in a graph. However, they simply treat any network as a homogeneous one as ignoring node types and relation labels. The same problem occurs in network statistics analysis [13] and community detection [4][8][14] for social networks.

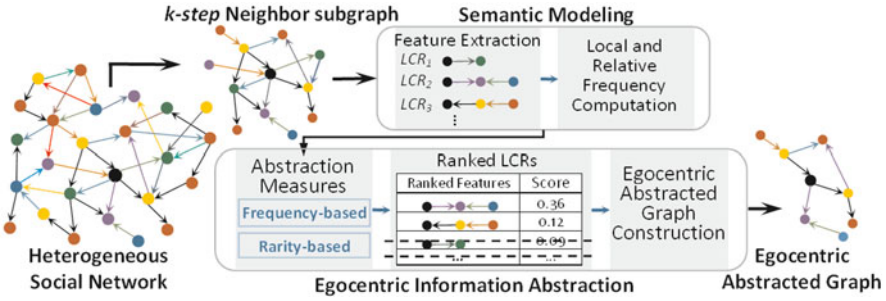


Fig. 3: Flowchart of proposed egocentric abstraction.

To handle the above issues and provide an intuitive, unsupervised, and efficient mechanism for egocentric information abstraction, we propose a model integrating both symbolic and statistic retrieval techniques. The flowchart is shown in Figure 3. We model the semantic behaviors of the ego using the surrounding substructure of the k-step neighbor subgraph. The Ego-based features (linear combination of relations) are extracted to represent the entities, and the corresponding feature values are calculated using sampling techniques on nodes and paths. Moreover, we propose three abstraction measures, namely local frequency, local rarity and relative frequency, to serve as the distilling criteria to perform abstraction from diverse views. Finally, we construct the abstracted graph for visualization using the distilled information. Our contributions and advantages are listed as follows:

1. We define and propose the solution for discovering representative egocentric abstraction in heterogeneous social networks to facilitate advanced analysis and visualization on social networks.
2. Both topological and relational (i.e., semantic) information are integrated as the linear combination of relations in our model to capture the behaviors of the given ego node. Besides, we propose the linear and cyclic features to describe the ego node.
3. Three abstraction views are introduced, and each of which encompasses its own physical meaning.
4. We conduct experiments on both natural and synthetic datasets to demonstrate the validity and usefulness of our system. The results indicate that our abstraction mechanism can indeed capture certain important information of the ego and provide accurate and efficient solutions for crime identification.

This paper is organized as follows. We describe the related work in Section 2 and the methodology is presented in Section 3. Section 4 reports the experiments. We discuss some relevant issues in Section 5 and conclude in Section 6.

2 Related Works

We divided the existing works related to information abstraction on network data into the following four categories:

Graph Summarization. Graph summarization is about generating the compact summarized representation for a large graph. L. Zou et al. [16] propose summarizing a graph using the topological information of the original homogeneous graph. It is not a trivial matter questioning how their approach can be adopted to heterogeneous graphs. Y. Tian et al. [12] introduce the OLAP-style operations to summarize multi-relational graphs, in which users can apply drill-down and roll-up to control summarized resolutions. However, they only use the immediate links of nodes and the high-order relationship information is ignored. S. Navlakha et al. [7] use the principle of Minimum-Description-Length to summarize single-relational graphs. They allow lossless and lossy graph compressions with bounds on the indicated error, and produce the aggregate graph. Nevertheless, it is not clear how their method can be applied to a heterogeneous network.

Network Abstraction for Visual Analysis. Network visualization aims at efficiently displaying a large network by drawing the structural data with some simple analyses for human explorations. P. Appan et al. [1] summarize key activity patterns of social networks in the temporal domain using a ring-based fashion. L. Singh et al. [11] develop a visual mining program to help people understand the entire multi-mode networks at different abstraction levels, in which the abstraction is performed by merging or dividing among different types of entities. Z. Shen et al. [10] divide abstraction into structural and semantic parts, and present a visual analytics tool, OntoVis, where the relations in heterogeneous networks are reduced based on the concept of network ontology. However, all three suffer from insufficiently providing egocentric views to facilitate explorations. Besides, they consider simply links in the one step neighborhood of each node. We argue that high-order topological and relational information should be modeled to produce more meaningful abstraction from diverse aspects through combining our existing methods [17] with the proposed signature profiles.

Network Skeleton. Network skeleton refers to the hidden structural backbone of the network in a macro view. They preserve various topological properties of the graph, and thus can be regarded as a kind of abstraction from the global perspective. A. Y. Wu et al. [18] use recursive graph simplification to construct a multilevel mesh, which is a reduced graph of microclusters and

preserves the characteristics of scale-free networks. D. Vincent and B. Cecile [19] perform transitive reduction on directed graph data, which is an edge-removing operation aiming at retaining the reachability between nodes. They define transitive reduction as a minimal subgraph with the same transitive closure as the original graph. By detecting the overlapping maximal cliques as supernodes, N. Du et al. [20] propose to create the backbone graph of the supernodes using the minimum spanning tree algorithm. Though the network skeleton approaches can simplify the network to some extent, it is unclear how their methods can be adopted to incorporate heterogeneous information.

Mining in Heterogeneous Networks. While most existing social network analysis studies concentrate on the homogeneous networks, some efforts are gradually shifted to the heterogeneous networks recently. D. Cai et al. [3] try to detect the community structures based on user-specified relations with importance weighting in heterogeneous social networks. They tackle this problem through learning an optimal linear combination for user-given relations to find the most relevant heterogeneous network structures. J. Zhang et al. [15] consider the importance of entities and relations to recommend objects for users in a multiple layer information network. They propose a pair-wise entity learning algorithm and integrate a modified random walk mechanism to devise the recommendation method. S. Lin et al. [6] propose an unsupervised mechanism to model the heterogeneous information surrounded each entity to identify the abnormal instances and generate reasonable explanations for them in a multi-relational social network.

3 Methodology

The formal definition for egocentric information abstraction in a heterogeneous social network is given as follows.

Given: *(a) a heterogeneous social network H, (b) the query vertex x representing the ego, and (c) the information filtering threshold* $\delta(\leq \delta \leq 1)$ *to control the level of abstraction.*

Outputs: three egocentric abstracted graphs of x, each of which belongs to the subgraph of H and corresponds to one of the three proposed abstraction views, as described in 3.3.

Definition 1. Heterogeneous Social Network). A heterogeneous network H(V, E, L) is a directed labeled graph, where V is a finite set of nodes, L is a finite set of labels, and $E \subseteq V \times L \times V$ is a finite set of edges. Given a triple representing an edge, the source, label, and target map it onto its start vertex, label, and end vertex, respectively. The function types$(V) \rightarrow r_1, ..., r_j, r_i \in L, j \geq 1$ maps each vertex onto its set of type labels.

A heterogeneous social network consists of the topological part and relational part. The nodes are various types of actors, each of which is surrounded

by certain combinations of diverse links and nodes. Here we propose to summarize the semantics of a given ego node via combining its surrounding linear substructure together with the statistical dependency measures obtained through certain sampling techniques. The egocentric information abstraction contains four main stages. First, a set of features, including linear and cyclic features, are automatically selected and extracted based on the surrounding network substructure of the given ego node. They will serve as the basis of summarization. Second, the statistic dependency measures between the features and the ego node are generated. Third, we apply certain distilling criteria to remove less relevant information. Finally, an egocentric abstracted graph can be constructed in an incremental manner that allows the users to visualize the results. The elaboration of these four stages is provided in section 3.1 to 3.4.

3.1 Ego-based Feature Extraction

We first extract the k-step neighbor subgraph $H_{k,x}$ of the ego node x. Constraining on the size of the neighborhood is reasonable since it is usually assumed farer away nodes do not have as significant inference as closer ones do. Then we propose to extract the *linear combination* of relations (LCR) as the base to represent the surrounding structure of the ego node. A LCR is defined as an ordered sequence of relations starting from the ego x. The linear combination of relations can be exploited to capture different kinds of features as behaviors for the ego x. Here we divide the features of x to two categories based on the characteristic of LCR_s: linear and cyclic features of LCR_s, as shown in Figure 4. Linear features are simply relational paths starting from the given ego x to one other vertex in the network. The linear features can be exploited to capture the interaction between x and its neighbors. Cyclic features represent paths that form a cycle. As shown in Figure 4, we consider three kinds of relational structures: self loop, triangle, and quadrangle, and each of which possesses certain physical meaning. Self loop implies a given node x has multiple and potentially diverse interactions with the other node. Triangle cycle can be regarded as a sign of highly impact triple that captures the three-way relationship among objects. Finally we exploit the quadrangle structure of LCR_s to highlight the intermediate mediators between nodes. We employ the cyclic features only up to quadrangle due to the computation complexity as well as the lack of physical meanings for those higher-order cycles. Note that it is possible that some nodes in the network do not contain some of the three cyclic features.

For example, assuming the path length k=2, the set of distinct LCR_s of node A_1 in Figure 2 is shown in Table 1. Each LCR can be regarded as a kind of behavior of A_1. Note that the inverse edge set E^{-1} is the set of all edges $(v_1, \wr^{-1}, v_2)$ such that$(v_2, \wr, v_1) \in E$. And we only regard the direction of edges

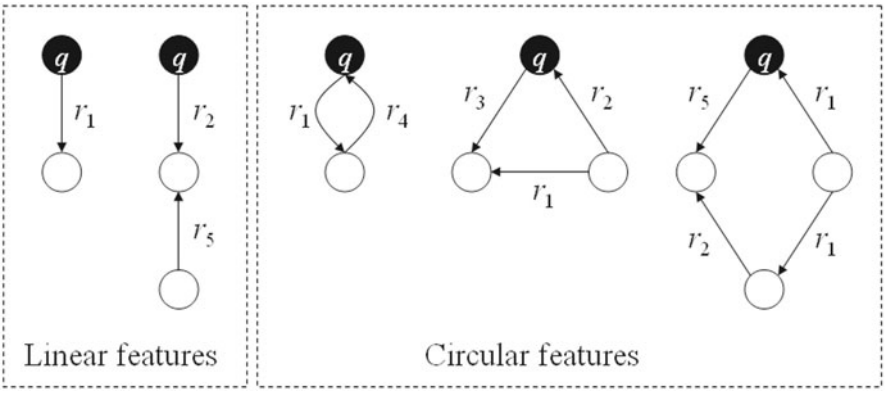


Fig. 4: The features of linear and cyclic relational structures.

Table 1: Two-steps LCRs from A1 of Figure 2.

Linear Features	LCR_1	$< hasActor^1, writeScript^1 >$
	LCR_2	$< hasActor^1, direct^1 >$
	LCR_3	$< hasActor^1, direct^1 >$
	LCR_4	$< hasActor^1, hasActor >$
	LCR_5	$< hasActor^1, originateFrom >$
	LCR_6	$< produce, direct >$
Cyclic Features	Self Loop LCR_7	$< produce, hasActor >$
	Triangle LCR_8	$< spouseOf, writeScript, hasActor >$
	Quadrangle LCR_9	$< produce, direct, direct, hasActor >$
	Quadrangle LCR_10	$< hasActor, direct, direct, hasActor >$

for linear features. For cyclic features, we do not consider the information of inverse and simply record the relations. Besides, we only record once for each cyclic LCR.

3.2 Nodes and Paths Sampling

In this section we perform certain statistic sampling on these extracted linear combination of relations (i.e., ego features) to compute the feature values. Two independent and identically-distributed (I.I.D.) random experiments are designed and applied. In the first random experiment (RE_1), we randomly select a node x from the network, then randomly select an edge e_1 starting from x, denoted by $< x, e_1, y >$, further randomly select another edge e_2 starting from y, denoted by $< y, e_2, z >$, and so forth. This stops when the number of edges chosen reaches k. The second one (RE_2) looks very similar to the first, except that we start from a randomly chosen edge $< a, e, b >$

instead of a node. Next we randomly pick another edge starting from node b. Again, this continues until k edges are chosen. The outcomes of either experiment is a path, and which we can define two random variables X and L. X represents the starting node of that path and L represents the LCR of this path. Note in this example, an instance of X is represented as x and one instance of L is $< e_1, e_2, ..., e_k >$. We use $X1$ and $X2$ to denote the starting node produced by RE_1 and RE_2, and the same for LCR_1 and LCR_2. With these four random variables, we then define two conditional probability mass functions: $P(L_1 = \wr | X_1 = x)$ and $P(X2 = x | L2 = \wr)$. We call the former *local frequency* of the ego node x, since it essentially stands for the probability that the LCR of a randomly picked path from x in face equals $\wr$. On the contrary, we call the latter relative frequency of an ego node since it represents the probability that an ego x is involved as the starting node in a given LCR $\wr$. The former is called "local" because this particular LCR is compared with other LCR_s starting from the same ego node (regardless how it distributes in the rest of the network). The latter is called "relative" or "global" since its value depends on how it is distributed in the entire network.

Table 2: Conditional Probabilities of $RE_1 : P(L_1|X_1)$. ($tb\wr_{local}$)

	LCR_1	LCR_2	LCR_3	LCR_4	LCR_5	LCR_6	LCR_7
x_1	0.02	0.08	0	0	0.1	0.3	0.5
x_2	0.3	0.03	0.4	0.25	0	0	0.02
...	...	...	...	...	...	...	...
x_{100}	0	0	0.01	0.07	0.9	0	0.02

After sampling both RE1 and RE2 for sufficient amount of times, it is possible to create two tables: $tb\wr_{local}$ and $tb\wr_{realtive}$ (e.g. probability values in Table 2 and 3, assuming there are only 7 LCR_s) which consist of the corresponding conditional probabilities. We call such tables the vector-based summarization of nodes. That is, each row vector in the table is a summarization of one node in the network. Note that in Table 3 we also show the rank (i.e., comparing with all nodes of the same type) of each $P(X_2|L_2)$ below its value inside the parentheses. Besides, the probability of each row sums to 1 in Table 2 while in Table 3 the probability of each column sums to 1.

3.3 Information Distilling

We propose two policies, frequency-based and rarity-based, to distill information from different views. Rarity and frequency basically occupy two opposite

Table 3: Conditional Probabilities of $RE_2 : P(X_2|L_2))$. $(tbl_{relative})$

	LCR_1	LCR_2	LCR_3	LCR_4	LCR_5	LCR_6	LCR_7
x_1	0.05 (76)	0.15 (5)	0.31 (2)	0 (99)	0.06 (88)	0.28 (3)	0.01 (34)
x_2	0.15 (22)	0 (66)	0 (72)	0.7 (1)	0.09 (32)	0.01 (68)	0.08 (21)
...	...	...	...	...	...	...	...
x_{100}	0 (82)	0.01 (60)	0.56 (1)	0.05 (38)	0 (93)	0.02 (51)	0.12 (12)

ends of the spectrum, and each reveals either important or meaningful information about the ego. Frequent behaviors are generally important for pattern recognition and rare events can sometimes lead to certain novel discoveries. Combining the two views (i.e., local and relative view) and two policies (i.e., frequency-based and rarity-based), four abstraction measures can be created, as shown in Table 4. Here we abandon the relative rarity view since it does not possess an apparent real-world meaning. Below we illustrate the ideas of the rest three views via an example using the above two tables.

Table 4: The four abstraction measures from two viewpoints.

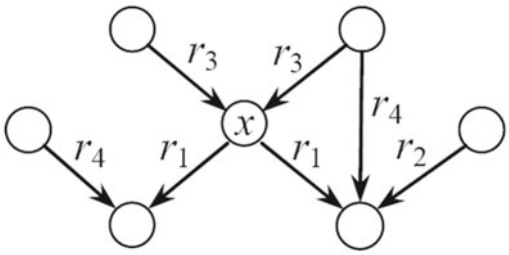
	Local	Relative
Frequency	Local Frequency	Relative Frequency
Rarity	Local Rarity	Relative Rarity

1. Local Frequency. It chooses the frequent $P(L_1|x)LCR_s$ from the vector of the given ego node x as the important ones. For example, if the threshold δ is set to $2/7$, only the top two frequent LCR_s in Table 2 $(i.e., LCR_6 and LCR_7)$ are picked to represent x_1. In other words, LCR_1 to LCR_5 are filtered out since they do not occur as frequent as other LCR_s with respect to x. The idea behind this view is that x is summarized by the frequent behaviors it involves.
2. Local Rarity. Opposite to local frequency, the rarity view of abstraction keeps the rare events happening to x and ignores the frequent ones. For the sample example δ is set to $2/7$, LCR_1 and LCR_2 will be distilled while the rest will be ruled out. Note that the "rare events" consider only those happening at least once, therefore excluding LCR_s whose conditional probabilities are zero such as LCR_3 and LCR_4. The idea behind this view is that the rare LCR_s could indicate something that should not happen but

in fact still occurs, and thus demands more attention. The other reason such view of abstraction should exist is that the rare events in a large network are generally harder to be detected by human beings than the frequent ones.

3. Relative Frequency. This uses Table 3 instead of Table 2. $P(X_2 = x|L_2 = \wr)$ in fact represents how frequent the ego x is involved in $\wr$ compared to other nodes. Since $\sum_X P(X_2|L_2) = 1$, we can treat each column in Table 3 as a relative comparison among all nodes for a certain LCR $\wr$. Then $P(X_2 = x|L2 = \wr)$ is representative of x if this value is relatively high compared to other nodes. In the example, LCR_3 and LCR_6 will be chosen to represent x since they are relatively high (i.e., ranked 2^{nd} and 9^{th}) compared to other nodes. The idea behind this view is that it picks the features best distinguishing x from others. Furthermore, since a heterogeneous social network generally has different types of nodes, it makes more sense to only compare the nodes of the same type when determining the rank of $P(X_2|L_2)$. For instance, it might not make sense to compare the number of publications among people from different research areas.

Note that the abstraction measures of information distilling are applied to LCRs of linear and cyclic features independently since they carry different kind of information and shall not be put on the equal ground for comparison, and usually the linear features occur more frequently than the cyclic ones since in the former case the target node is not constrained to be equivalent to the source.



(a)

ID	Ranked LCRs	Score
rs_1	$\langle r_1, r_4^{-1} \rangle$	0.36 (2)
rs_2	$\langle r_1, r_2^{-1} \rangle$	0.08 (5)
rs_3	$\langle r_3^{-1} \rangle$	0.09 (10)
rs_4	$\langle r_3^{-1}, r_4 \rangle$	0.02 (79)
rs_5	$\langle r_1 \rangle$	0.005 (99)

(b)

Fig. 5: (a) An Example $H_{k,x}$ and (b) the ranked LCR_s.

3.4 Abstracted Graph Construction

Now we have distilled features as the abstraction for an ego node. One plausible form is to report distilled LCRs and corresponding probabilities to the users. Though it seems to be a reasonable output since $P(L_1|X_1) or P(X_2|L_2)$ can serve as a term that explains why such an abstraction is made, an alternative and more understandable way is to convert the distilled information back to a graph. Here we use an incremental method to obtain a subgraph composed of only distilled LCRs and the corresponding nodes.

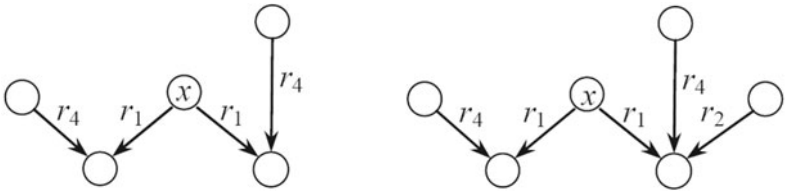


Fig. 6: The Abstracted graph after adding LCR_1 and (b) the final graph after LCR_1 and LCR_2 are added.

Figure 5 and 6 illustrate this idea. Assume we want to keep the top two scored LCRs and filter out the rest. The LCR with the highest score (i.e., LCR1) is first used to match the original network to obtain a subgraph that originates from the ego x and contains all nodes and edges involved in LCR1 (see Figure 6(a)). The same action is performed for LCR2. The final egocentric abstraction of x is shown in Figure 6(b). Note that it is not feasible to produce the abstracted graph by removing the discarded LCRs since edges involved in one LCR might also occur in others. Therefore, eliminating one of them will sometimes cause the informative LCRs to disappear. The complete algorithm of our egocentric information abstraction is given in algorithm 1. We first extract the k-step neighbor subgraph for the given ego node (line 1), and then perform sampling to derive local and relative tables (line 3-9). According to the three designated viewpoint of abstraction, the most relevant linear combinations of relations are picked (line 10-18). Finally, the abstracted graph is constructed incrementally (line 19-23).

4 Evaluations

We perform two experiments. The first focuses on demonstrating how the proposed framework can be performed on a real-world movie network by showing the resulting abstracted graph based on the proposed features using

Algorithm 1 Egocentric Information Abstraction

Input: H: a heterogeneous network; x: the query ego node; k: the step size for linear combination of relations; δ: the information filtering threshold; *view*: policy for information distilling; *feature_option*: option for the linear and cyclic features.
Output: H^{abs}: the abstracted graph from different views

Extract the k-step neighbor subgraph $H_{k,x}$ of x.
LCR = retrieve LCRs for *feature_option* from x.
derive the table of local measure
$tbl_{local} = P(L_1|X_1)$ using LCR.
derive the table of relative measure and rank each column
$tbl_{relative} = P(X_2|L_2)$ using SP.
for j = 1 to $|LCR_s|$ **do**
    Compute the ranked value of $tbl_{relative}(:, j)$ in descending order.
end for
distilledSet = { }. // collect the LCRs of top score of specified view
if view = "localFrequency" **then**
    $distilledSet = distilledSet \cap argmaxOfTop\delta(tbl_{local}(x, LCR_i))$.
else if view = "localRarity" **then**
    // note that those scores equal to zero are ignored
    $distilledSet = distilledSet \cap argminOfTop\delta(tbl_{local}(x, LCR_i))$.
else if view = "relativeFrequency" **then**
    $distilledSet = distilledSet \cap argmaxOfTop\delta(tbl_{local}(x, LCR_i))$.
end if
Let H^{abs} = NULL.
for cr *distilledSet* **do**
    instances = Find path instances in $H_{k,x}$, whose LCR equals to cr.
    $H^{abs} = H^{abs} \cap$ instances.
end for
return : H^{abs}

different abstraction measures. The second experiment is designed to assess the quality of the abstraction through human studies on a crime dataset. The goal is to find out whether the egocentric abstraction can improve the accuracy and efficiency of human decisions.

4.1 Case Study for a Movie Network

We apply our egocentric information abstraction on a movie dataset to exhibit the abstracted graphs via different abstraction views using linear and cyclic features respectively. The UCI KDD movie dataset [5] is used to construct the heterogeneous network. It contains about 24,000 nodes (9,097 movies, 3,233 directors, 10,917 actors, and some other movie-related persons such as producers and writers) and 126,926 relations. There are 44 different relation types which can be divided into three groups: relations between people (e.g. spouse and mentor), between movies (e.g. remake), and between a person

and a movie (e.g. director and actor), which makes it very difficult for human to analyze. Here we take "Meg Ryan", a famous actress, as the ego node to demonstrate the egocentric abstracted graphs for linear features of LCR. Also we choose "Tom Cruise", a famous actor, as another ego to present the egocentric abstracted graphs for cyclic features of LCR because this ego has more cyclic features. Besides, we would like to point out that this UCI KDD dataset is incomplete where some information is missing. Therefore certain statistics collected based on it might not reflect the real-world situation. The 2-step neighbor subgraph of "Meg Ryan" is shown in Figure 1. This is not a trivial network analysis since there are 116 nodes, 137 edges and 18 different LCRs.

Using the linear features, the abstracted graph of local frequency for Meg Ryan is shown in Figure 7, which captures the regular behavior of her. The filtering threshold $\delta = 20\%$ is applied in our abstraction (which implies we only keep 20% of the LCRs). We can observe that she played in many movies, especially in comedic, dramatic, and romantic categories. Besides, her husband, Dennis Quaid, is also an actor of many movies. They co-starred in three of them. Such information is not as trivial to obtain from the original graph.

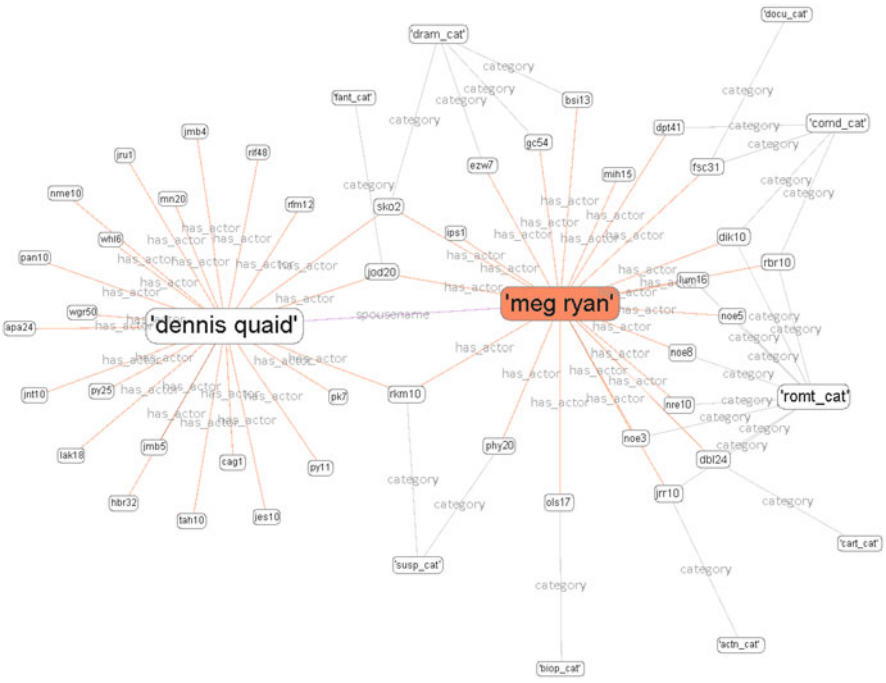


Fig. 7: Local frequency of "Meg Ryan".

Using the linear features, the local rarity view for Meg Ryan is shown in Figure 8. It captures the rare behavior of Ryan. We can observe she is also a producer of a movie (i.e., lak16). Besides, her husband's brother (i.e., Randy Quaid) also works in the movie industry (note that only movie-related persons are listed in this dataset). Finally there is a movie she acted (i.e., noe3) whose cinematographer (denoted as 'c' here) is listed in this dataset. This becomes a rare pattern for her since none of her other movies has such information recorded.

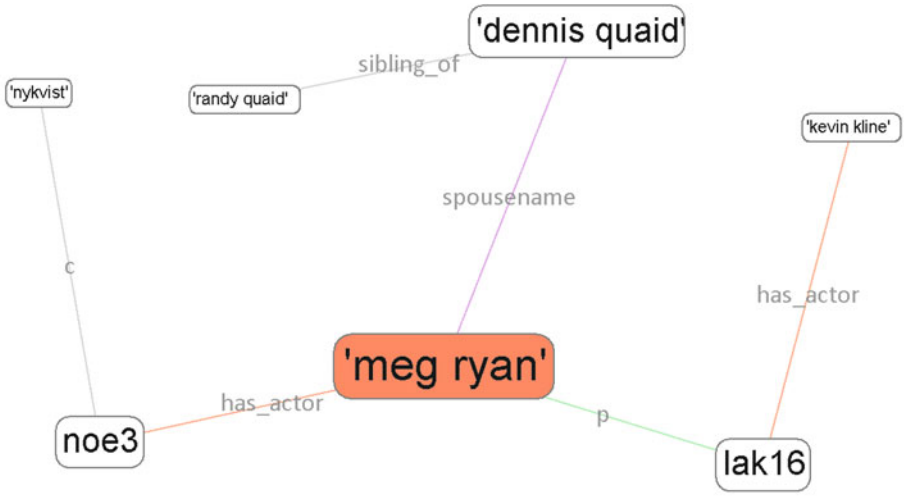


Fig. 8: Local rarity of "Meg Ryan".

Using the linear features, the abstracted graph of relative frequency for Meg Ryan is shown in Figure 9, which compares the behavior of Meg Ryan with other actors (note: not all other persons in the dataset) and identifies the behavior she significantly involves in. We can observe an interesting behavior of her as she acted in relatively large amount of remade movies comparing with others. Also she produced a movie (i.e., lak16) and such behavior does not appear to be frequent among other actors. Finally, one path of her based on local rarity measure, namely his husband's sibling is also a movie person, turns out to be rare among other actors as well, and thus becomes a relatively frequent behavior of her (that is, there are very few others in this dataset whose husband's sibling is also a movie person).

Using the cyclic features, the abstracted graph of local frequency for Tom Cruise is shown in Figure 10. We can observe a set of frequent quadrangle cycles, namely $< hasActor, category, category, hasActor >$, $< hasActor, hasActor, hasActor, hasActor >$, $< hasActor, p, p, hasActor >$, *and* $< hasActor, d, d, hasActor >$. The former two shows that Cruise acted in many movies of identical categories, and collaborated with many actors in different

Fig. 9: Relative frequency of "Meg Ryan".

movies. The latter two indicates that he frequently acted in movies directed by the same person, and frequently acted in movies that has the same producer. Besides, sometimes he preferred to incorporate with permanent directors and producers (e.g., the movie "tos5" and "tos10"). Using the cyclic features, the abstracted graph of local rarity for Tom Cruise is shown in Figure 11. First, he produced and acted in one movie (i.e., "bdp30"), which is not as frequent for him. Second, there are two cyclic structures of quadrangle: (1) Tom Cruise played in the same movie with his spouse's sibling (i.e., "Paul Abbott"), (2) it is quite rare for him to involve in two movies which is produced and directed by the same person (i.e., C. Crowe).

Using the cyclic features, the abstracted graph of relative frequency for Tom Cruise is shown in Figure 12. This shows that comparing with other actors in the network; it is quite frequent for him to act in the same movie with his spouse (i.e., Nicole Kidman). Besides, the movie "bdp30" appears again to reveal that in this dataset not many actors acted in a self-produced movie.

In this case study, we have used a heterogeneous movie network to demonstrate which kinds of information can be revealed through which egocentric view. We have also demonstrated that through our abstraction mechanism, it is possible to discover not only some expected details (e.g. Ryan acted in many romantic movies) but also some unexpected yet potentially interesting facts (e.g. Ryan acted in many remade movies and produced a movie) about

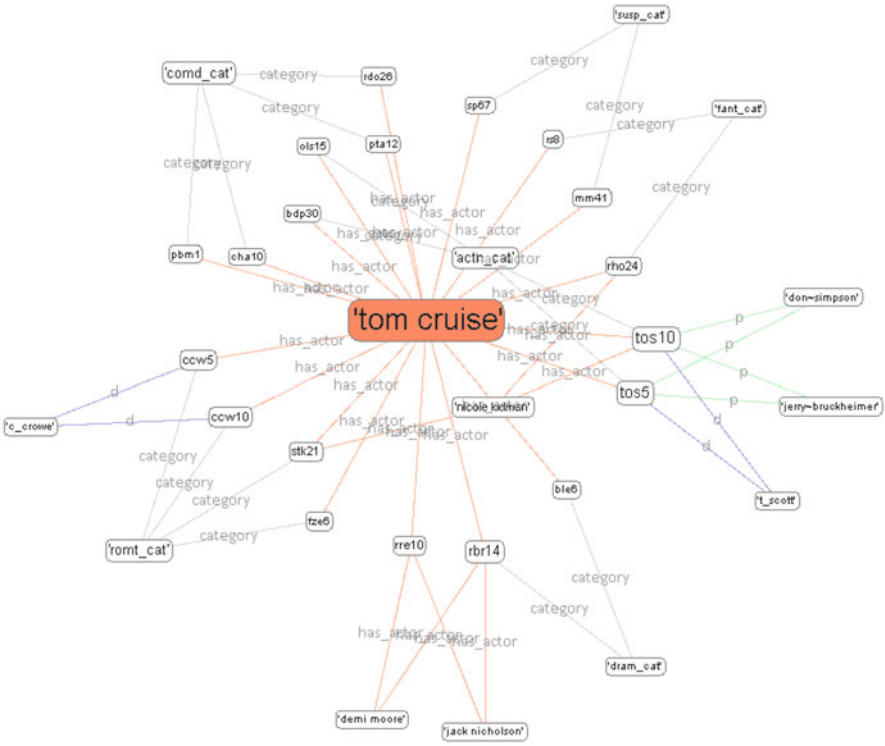


Fig. 10: Local frequency of "Tom Cruise."

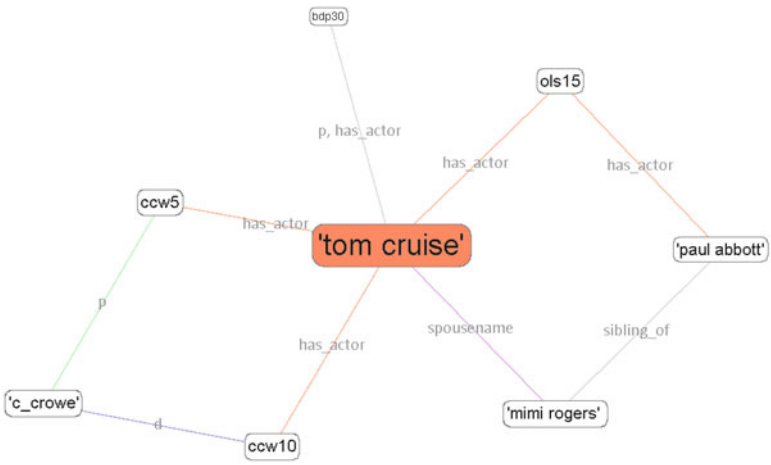


Fig. 11: Local Rarity of "Tom Cruise."

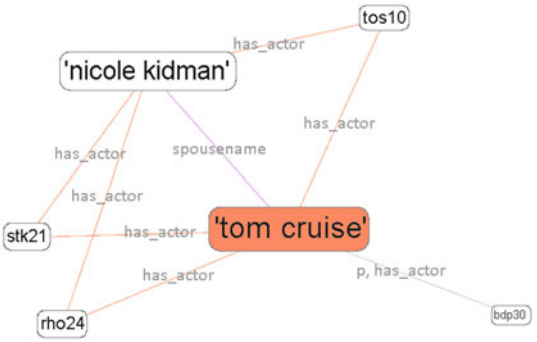


Fig. 12: Relative frequency "Tom Cruise."[5]

the ego node. It might even satisfy some hard-core fans by revealing certain information about her ex-husband.

4.2 Human Study for Crime Identification

In this experiment, we evaluate our abstractions through a study of the quality of human decision-making. The goal of this evaluation is three-fold: 1) to know whether and which of the abstracted networks can assist human subjects to make more accurate decisions. 2) To see whether the abstractions can reduce the time needed to make a decision. 3) To learn whether the abstraction can improve human subjects' confidence about their decisions.

The dataset we used is a simulated crime dataset developed during US Defense Advanced Research Projects Agency's Evidence Extraction and Link Discovery Program [9] for evaluating link discovery algorithms like group detectors, pattern matchers, and etc. The data is generated by a simulator of Russian organized crime (i.e., Mafiya) that simulates the process of ordering, planning, and executing criminal activities such as murders or gang wars with many possible variations and records an incomplete and noisy picture of these activities in the files (e.g. financial transaction, phone call, email, somebody being observed at a location, somebody being killed by someone unknown, etc.). It has about 9000 nodes and twice as many edges with 16 node types of objects (e.g. bankAccount, Mafiya, and industry) and 31 different relation types (e.g. perpetrator and victim). There are 42 gang nodes and 20 contract murder events. Besides, it is noisy since some relations are missed or labeled incorrectly, which could cause difficulties for analysts.

The experiment setup is as follows: we first choose 10 plausible gang nodes among which three were truly involved in the highest level events(i.e., gang war and industry takeover). For each gang node, three different views of

egocentric abstracted graphs were generated. Together with the original k-neighborhood graph (we choose k=3 in this experiment), we will have four different set of networks (each contains 10 independent networks corresponding to 10 plausible gangs) presented to the subjects.. To avoid interference among different tasks, the IDs of all candidate gangs are randomly given for each task. These four sets of resulting graphs are shown to a total of 20 human subjects (they were not told in which order of datasets they should pursue) and the users were asked to select three (out of ten) nodes that are most likely to commit high-level crimes for each set. Therefore, we can examine how many candidates were picked correctly for each set. Before the experiment, the subjects were asked to study the background knowledge of this domain so they understood the meaning of each relation and the node types as well as the meaning of the events. The four generated graphs of one criminal node are illustrated in Figure 13 to 16, which are corresponding to the original 3-neighborhood graph, local frequency, local rarity, and relative frequency in order. Note that the filtering threshold δ is set to 0.2, which implies we only keep 20% of the LCRs during abstraction. The black nodes are nodes representing criminal candidates.

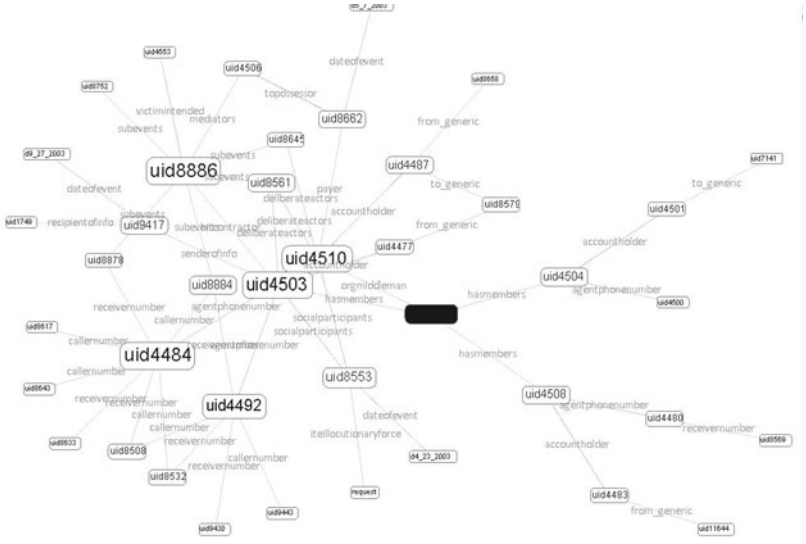


Fig. 13: The original 3-neighborhood graph.

The results are displayed in Table 5. We also show the improvement over k-step neighbor subgraph in the first column and 95% confidence interval for average time and confidence.

In terms of accuracy, the results show that users can usually perform better (the improvement can be as high as 13.3%) while using the abstracted

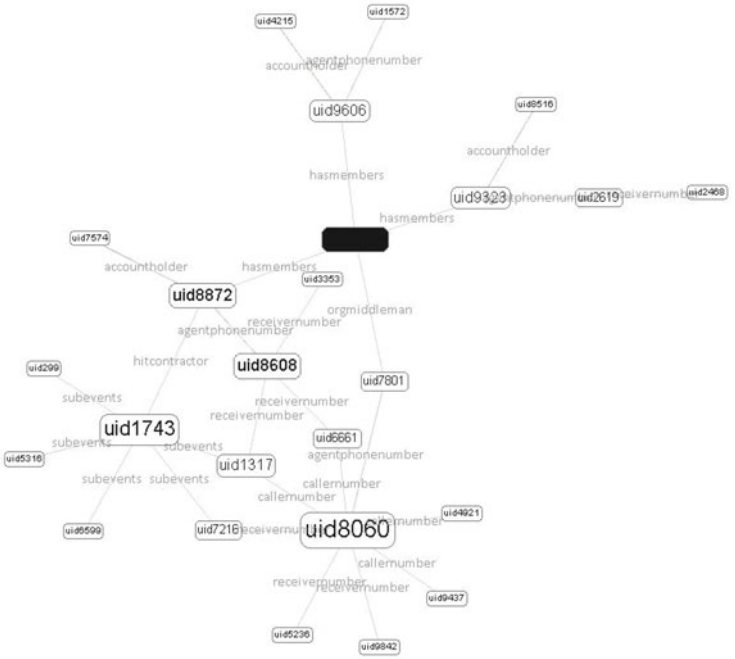


Fig. 14: Abstracted graph of local frequency.

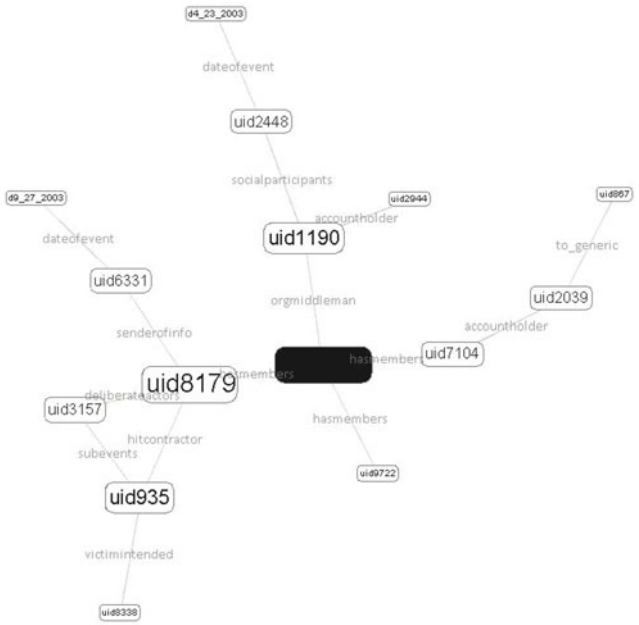


Fig. 15: Abstracted graph of local rarity.

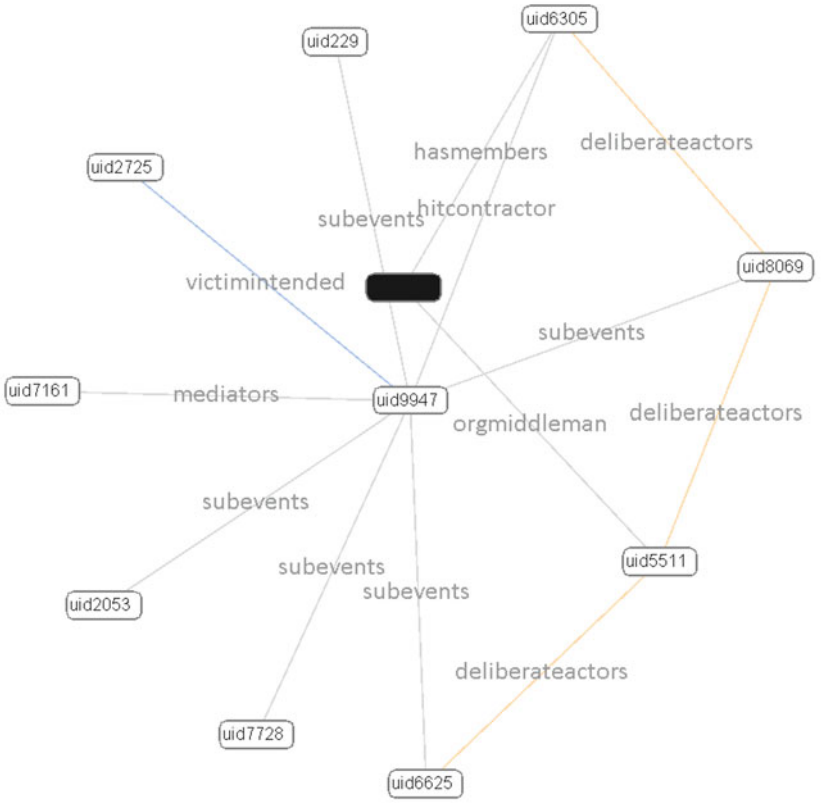


Fig. 16: Abstracted graph of relative frequency.

Table 5: Conditional Probabilities of $RE_2 : P(X_2|L_2))$. $(tbl_{relative})$

		Avg. Precision	Avg. Time (minutes)	Avg. Confidence (1∼5, 5 is the highest)
No Abstraction	k-neighborhood Graph	39/60	36.6 ± 6.6	3.15 ± 0.36
Using Abstraction	Local Frequency	41/60 (+3.3%)	18.9 ± 5.9	3.20 ± 0.35
Using Abstraction	Local Rarity	44/60 (+8.7%)	13.9 ± 3.7	3.45 ± 0.33
Using Abstraction	Relative Frequency	47/60 (+13.3%)	10.9 ± 2.2	3.73 ± 0.39

networks comparing with the original one. Our explanation is that although certain information is lost after abstraction, it is likely the critical messages are remained while some noise is filtered out, which leads to better results.

The major improvement, as shown in the second column of Table 4, lies in efficiency. Users spend significantly less amount of time ($< 50\%$) to reach better results. The improving on accuracy, efficiency, and confidence demonstrates that the abstraction is capable of facilitating better human analysis. In this dataset, there are some key evidences that can indicate the high-level events. After analyzing the abstracted graphs manually, we have realized that each abstraction view more or less captures different parts of those key evidences. For example, a kind of LCR that represents "the gang has hired some middleman intending to pursue something illegal" happens only to the high-level crime participants; therefore it can be highlighted using the relative frequency view, which becomes an important evidence for the human subjects to make the right decision. This could be the major reason that this view eventually leads to the best results among others.

5 Discussions

There are several issues worthy of further discussions:

1. The efficiency. To estimate the probabilities accurately, we need to sample a sufficient amount of paths, which becomes the bottleneck of our approach. However, a technique called likelihood weighting, which has been applied successfully in the inference procedure of Bayesian Networks, can be applied to force the occurrence of some rare events. Then the likelihood can be reweighted based on the frequency of the forced decisions.
2. Parameters. There are two parameters to control the level of abstraction: the k in k-neighborhood and as the trimming threshold. Each of them has its own physical meaning. Increasing k can enlarge the size (or radius) of the network and increasing can boost the density of the graph. Therefore we recommend determining k based on the number of nodes and links in the network, and adjusting based on the number of different link types.
3. Union or Intersect measures. In reality there can be more than three measures of abstraction since views can be integrated. For example, one can union local frequency and local rarity measures to visualize both frequent patterns and rare events in the abstraction. One can also intersect the local frequency and relative frequency views to make sure only behavior that is both frequent and representative are shown.

6 Conclusions

In this paper we present a method for egocentric information abstraction for heterogeneous social networks. We believe it can be applied to create a node-based search engine for social networks as well as realizing social net-

work visualization. Here we provide an alternative view about our approach. An intuitive approach to graph abstraction is to identify certain seems-to-be irrelevant edges and vertexes to remove. However, it is non-trivial how such removal can be made (either manually or automatically) when the information is represented as a heterogeneous social network where nodes and edges are mixed together to form complicate patterns. To answer this challenge, we argue that the abstraction should be pursued in a retaining manner rather than an *eliminating* manner. That is, we should build the abstracted graph by trying to keep important or relevant information instead of discarding the irrelevant ones. Therefore in this paper we propose a two-level abstraction schema. The first level of abstraction is to transform the original network into a vector-space representation using symbolic modeling and sampling techniques. The reason to perform such transformation is that now we are then allowed to pursue the second-level abstraction as applying some simple and intuitive criteria to determine which portion of the information should be retained. Finally our goal can be achieved through incrementally transforming the retained vectors back to the original domain of networks.

References

1. P. Appan, H. Sundaram and B. L. Tseng. Summarization and Visualization of Communication Patterns in a Large-Scale Social Network, In Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'06), 371-379, 2006.
2. S. Brin and L. Page. The Anatomy of Large-scale Hypertextual Web Search Engine. In Proc. of Intl. World Wide Web Conference (WWW'98), 107-117, 1998.
3. D. Cai, Z. Shao, X. He, X. Yan and J. Han. Mining Hidden Community in Heterogeneous Social Networks. In Proc. of ACM SIGKDD Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD'05), 58-65, 2005.
4. D. Chakrabarti and C. Faloutsos. Graph Mining: Laws, Generators, and Algorithms. ACM Computing Survey, 38(1), 2006.
5. S. Hettich and S. D. Bay. The UCI KDD Archive. http://kdd.ics.uci.edu, University of California, Irvine, Department of Information and Computer Science, 1999.
6. S. D. Lin and H Chalupsky. Discovering and Explaining Nodes in Semantic Graph. IEEE Transactions on Knowledge and Data Engineering, 20(8), 1039-1052, 2008.
7. S. Navlakha, R. Rastogi and N. Shrivastava. Graph Summarization with Bounded Error. In Proc. of ACM SIGMOD Intl. Conference on Management of Data (SIGMOD'08), 419-432, 2008.
8. M. E. J. Newman and M. Girvan. Finding and Evaluating Community Structure in Networks. Physics Review, 2004.
9. R. Schrag. A Performance Evaluation Laboratory for Automated Threat Detection Technologies. In Proc. of Performance Measures of Intelligent System Workshop (PerMIS'06), 2006.
10. Z. Shen, K. L. Ma and T. Eliassi-Rad. Visual Analysis of Large Heterogeneous Social Networks by Semantic and Structural Abstraction. IEEE Transactions on Visualization and Computer Graphics, 12(6), 1427-1439, 2006.
11. L. Singh, M. Beard, L. Getoor and M. B. Blake. Visual Mining of Multi-Modal Social Networks at Different Abstraction Levels. In Proc. of Intl. Conference on Information Visualization (IV'07), 672-679, 2007.

12. Y. Tian, R. A. Hankins and J. M. Patel. Efficient Aggregation for Graph Summarization. In Proc. of ACM SIGMOD Intl. Conference on Management of Data (SIGMOD'08), 567-580, 2008.
13. S. Wasserman and K. Faust. Social Network Analysis: Methods and Applications. Cambridge University Press, UK, 1994.
14. X. Xu, N. Yuruk, Z. Feng and T. A. J. Schweiger. SCAN: A Structural Clustering Algorithm for Networks. In Proc. of ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining (KDD'07), 824-833, 2007.
15. J. Zhang, J. Tang, B. Liang, Z. Yang, S. Wang, J. Zuo and J. Li. Recommendation over a Heterogeneous Social Network. In Proc. of Intl. Conference on Web-Age Information Management (WIAM'08), 309-316, 2008.
16. L. Zou, L. Chen, H. Zhang, Y. Li and Q. Lou. Summarization Graph Indexing: Beyond Frequent Structure-Based Approach. In Proc. of Intl. Conference on Database Systems for Advanced Applications, 141-155, 2008.
17. C. T. Li and S. D. Lin. Egocentric Information Abstraction for Heterogeneous Social Networks. In Prof. of Intl. Conference on Advances in Social Network Analysis and Mining (ASONAM'09), 255-260, 2009.
18. A. Y. Wu, M. Garland, and J. Han. 2004. Mining Scale-free Networks Using Geodesic Clustering. In Proc. of ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining (KDD'04), 719-724.
19. D. Vincent and B. Cecile. 2005. Transitive Reduction for Social Network Analysis and Visualization. In Proc. of IEEE/WIC/ACM Intl. Conference on Web Intelligence (WI'05), 128-131.
20. N. Du, B. Wu, and B. Wang. 2007. Backbone Discovery in Social Networks. In Proc. of IEEE/WIC/ACM Intl. Conference on Web Intelligence (WI'07), 100-103.

PROG: A Complementary Model to the Social Networks for Mining Forums

Anna Stavrianou, Julien Velcin, and Jean-Hugues Chauchat

Abstract Online discussion systems in the form of forums have been represented by graphs and analyzed through social network techniques. Each forum is regarded as a social network and it is modeled by a graph whose vertices represent forum participants. Here, we focus on the structure and the opinion content of the forum postings and we are looking at the social network that is developed from a semantics point of view. We formally define a model whose purpose is to provide complementary information to the knowledge extracted by the social network model. We present structure, opinion, temporal and topic-oriented measures that can be defined based on the new model, and we discuss how these measures facilitate the analysis of an online discussion. Applying our model to a real forum found on the Web shows the additional information that can be extracted.

1 Introduction

The abundance and popularity of online discussion systems that come in the form of forums, blogs or newsgroups, has pointed out the need for analyzing

Anna Stavrianou
ERIC Laboratoire - Université Lumière Lyon 2, Université de Lyon, France
e-mail: anna.stavrianou@univ-lyon2.fr

Julien Velcin
ERIC Laboratoire - Université Lumière Lyon 2, Université de Lyon, France
e-mail: julien.velcin@univ-lyon2.fr

Jean-Hugues Chauchat
ERIC Laboratoire - Université Lumière Lyon 2, Université de Lyon, France
e-mail: jean-hugues.chauchat@univ-lyon2.fr

and mining such systems. Monitoring how the users behave and interact with each other, their ideas and opinions on certain subjects, their preferences and general beliefs is significant.

Most existing works view an online discussion as a network in which users meet and contact each other, form communities and acquire certain roles. Forums are usually modeled by a graph whose vertices represent users that are connected with each other according to who speaks to whom. Such graphs are analyzed by social network techniques [2].

The application of the social network model to a forum provides information about how the users interact with each other. The structure of the discussion and the opinion information contained in the forum is lost. By taking into account the structure of the postings and their opinion content, we can become more familiar with the users and get to know better their attitude during the discussion. We can observe whether there is an important opinion presence in the forum and if so, we can measure its amount. In this way, it can be easily identified when users agree with each other, in which parts of the forum they are contradicted or whether they keep talking negatively.

In this article, we present a theoretical work that has been carried out with the purpose of looking at the social network developed in a forum from a semantics and opinion-oriented point of view. The contribution of our work is a new model which is complementary to the social network model. It can be applied to an online discussion together with the social network model in order to enrich the information extracted from a discussion. The proposed model goes beyond the exploitation of the developed user network. It emphasizes on the structure of the discussion and on the content from a topic- and opinion-oriented point of view. It combines Text and Opinion Mining techniques with Social Network Analysis concepts.

The rest of the article is structured as follows. Sect. 2 discusses related work. Sect. 3 presents the proposed model by defining its properties and presenting the components it is consisted of. In Sect. 4 we define structure, opinion, time and topic-oriented measures. In Sect. 5 we show how the proposed model can be applied to a real web forum and what information we can extract from it. Sect. 6 concludes by highlighting future perspectives.

2 Background

Most research regarding forum analysis focuses on analyzing the interaction between users or discovering how users form communities and are affected by them. Until now we have seen no works that examine automatically the presence and exchange of opinions inside a forum. The way in which opinions appear, influence and flow within a network of messages is not currently

discussed in the existing research. Nevertheless, our work has been influenced by the analysis presented in certain works in the social network domain.

One of these works is presented in [7] who analyze the Innovation Jam 2006 among IBM employees and external contributors. The representation of the discussion is seen from the point of view of postings rather than users. The difference from our work is that they do not consider the opinion flow inside the discussion. Their objective is to find out the degree of innovation of a discussion from a topic-oriented point of view and not to identify the opinion flow in it. Moreover, while in the IBM Innovation Jam the users are known, in our work users remain anonymous. Anonymous users tend to express more freely their opinions and, as a result, in the type of forums we analyze we can have more interesting opinion results.

Forum analysis has also been dealt with in [15]. They analyze the Java Forum by using Social Network Analysis methods for the purpose of automatically identifying user expertise. They represent the social network of the forum with a graph whose vertices represent users. Their objective is different from ours since we concentrate on the content rather than the participants of a discussion and we do not seek to find experts. Focusing on the content and the opinion exchange inside a discussion leads us to represent a discussion from the point of view of the content rather than that of users.

In [1] they attempt to separate a set of newsgroup users in those that are for or against a topic. They represent a newsgroup as a graph with user nodes and they base their analysis on the "reply-to" links between the users. Again, they focus on the users and not on the postings. Although they consider the presence of agreement and disagreement, they do not actually take the opinion into account.

Roles are assigned to user nodes of a graph in [4] and [11]. [4] analyze newsgroups by applying social network techniques and they interpret online communities by assigning roles to the members of the groups. This is done by observing how people relate to each other in a graph-based model of post-reply relations. They notice that short discussion threads point out question-answer exchanges and longer threads indicate proper discussions. [11] introduce a new measure that defines the number of communities to which a vertex is attached. Using this measure they assign roles to vertices by considering the community structure in the network of the vertex. Defining roles in this way, improves the performance of link-based classification and influence maximization tasks. We have been inspired by these works in the sense that each node is different and its position in the network carries information about how it affects the rest of the network. Both works, though, differ from ours both in the representation and the objective aspect.

From the existing works we can see that the current representations place the user in the center of the discussion with the interest in seeing how the user interacts and what his role as a discussion participant is or becomes. One main characteristic, though, of a discussion is the content. The users can take different roles according to how they express themselves, what opinion

they hold and how their opinion can be influenced and change throughout a discussion. These remarks lead us to consider a different representation for such discussions that exploits not only the interaction between users but also the exchange of information and opinion inside the developed network.

3 Post-reply Opinion Graph

In this Section we present a framework which achieves a forum representation complementary to the existing techniques. The new representation allows us to exploit the structural characteristics of a forum and analyze it from a semantics-oriented point of view. The proposed framework represents a forum by a "Post-Reply Opinion Graph", whose definition follows.

Definition 1. A **Post-reply Opinion Graph** (PROG) is a directed graph $G = (V, E)$ with a vertex set V and an edge set E. Each vertex represents a *posting* and each edge $e_{xy} = (x, y)$ points out a reply direction from the vertex x to the vertex y. A *posting* v represented by a vertex is defined as:

$$v = (m_v,\, op_v,\, u_v,\, tm_v),$$

where m_v is the actual content of the message, op_v the opinion polarity included in the message, u_v the user that has written it, and tm_v the timestamp that shows when the message was posted.

In Fig. 1 we can see an example of a Post-Reply Opinion Graph.

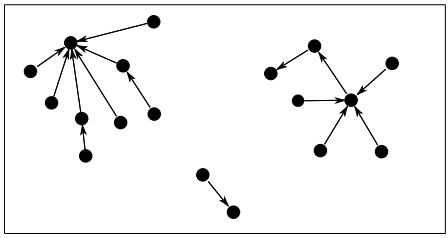


Fig. 1: A Post-Reply Opinion Graph of an online discussion

The graph has initially been presented as an opinion-oriented graph in [13]. At that moment the opinion polarities were used as labels on the edges of the graph. We have found out that having the polarities defined locally as part of the vertex allows us to extract easier and more efficiently information.

One main characteristic present in the definition of a Post-Reply Opinion Graph is the *opinion* op_v of a vertex v of the graph. The *opinion* $op_v \in$

$\{n, o, p\}$ captures the opinion polarity expressed in the message m_v. It may be negative (n), positive (p) or objective (o), when there is no opinion included. The opinion polarity is calculated by Opinion Mining techniques such as those presented in [3, 5, 6, 9, 14].

The *author* of the message u_v is encapsulated in the message object. In this way, information about the author is not lost. As a result, the social network of users can be extracted from the proposed model. This is an important property of the Post-Reply Opinion Graph, since the information provided by the social networks can still be exploited.

The notion of time is also encapsulated in the proposed model, so the future and the past of a vertex can be easily traced. The successor of a node v, which is unique in the case of PROG models, is a message object that has taken place immediately before the message object represented by the node v. Similarly, the predecessors of the node v, $\{v' \in V : (v', v) \in E\}$ contain message objects that have been posted after the posting represented by v.

We can understand this concept through a small example: let us assume that in a certain discussion, the posting v_3 sent at time tm_3 replies to the posting v_1 posted at time tm_1. In the same discussion, the posting v_2 posted at tm_2 replies to no message and it was sent after the posting v_1 and before the v_3 i.e. $tm_1 < tm_2 < tm_3$.

The posting v_2 may be a message that has actually been influenced by existing messages but does not explicitly reply to one of them. As it can be seen from Fig. 2, the graph can be enriched with the knowledge of the chronology. On the right hand side of the Fig., the chronology edges are shown with dotted lines. This can help us to better identify which postings may have influenced a message not just by exploring the structure of the discussion, but also by analyzing the chronology links. In the example shown in Fig. 2, the author of v_3 has replied to v_1, but it may be after s/he has read the posting v_2. Exploiting the chronology links is a future issue in the Post-Reply Opinion Graphs.

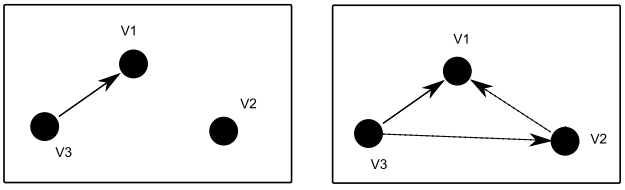


Fig. 2: The chronology knowledge enriches the structure of a discussion. In the Fig., v_2 has been sent after v_1 and before v_3, while v_3 replies directly to v_1

Apart from the implicit temporal information, the proposed graph includes the information of time explicitly in each vertex. The chronology of each

posting is captured through the tm_v timestamp. Knowing the time of having posted a message allows us to enrich the structure of the graph.

3.1 Properties

Let us consider a PROG graph $G = (V, E)$ and one of its vertices $v \in V$. The number of vertices adjacent from that vertex v is called the *outDegree* of v. According to the theory of graphs, the *outDegree* of a vertex v is defined as:

$$outDegree(v) =\mid \{v' \in V : (v, v') \in E\} \mid . \tag{1}$$

Lemma 1. *The outDegree of the vertices of a PROG graph G is 0 or 1.*

This lemma stands because by definition, in the online discussions we model, the user can post a message that replies to maximum one existing posting.

Another characteristic of PROG graphs is that a *walk* between two vertices does not repeat any edge or vertex. Thus, we have the following theorem:

Theorem 1. *In a PROG graph G, all possible walks between two vertices are paths.*

Proof. Suppose the PROG graph G has a walk which is not a path. This means that there is a walk that repeats at least one vertex of the graph. Let this repeated vertex be v. In order for this vertex to be repeated in the walk, one of its "past" vertices v' should be adjacent to it so that an edge $e_{v'v} = (v', v)$ exists.

Since the vertex v' has happened before v, we have $tm_v > tm_{v'}$. Thus, there cannot be an edge $e_{v'v}$ because the posting represented by v' cannot reply to the posting represented by v since v happened after v'. Hence, no such walk exists that is not a path. □

Product of the previous theorem is the insight that a PROG graph cannot contain cycles. Thus, it is a forest since it contains no cycles and it can be consisted of many components. Therefore, we have the following Lemma:

Lemma 2. *A Post-Reply Opinion Graph is a forest.*

3.2 Components

A PROG graph is consisted of components whose identification allows us to define measures in order to extract useful information from such graphs. Here,

we present two basic components; the discussion threads and the discussion chains.

Definition 2. The set of the **discussion threads** in a Post-Reply Opinion Graph G is the union of all the maximal connected components of G.

The discussion threads can be "queried" either by a message m or a user u. For example, the threads where the user u has participated can be found by tracing the vertices of each thread of the graph until a message object $v = (m_v, op_v, u, tm_v)$ is found.

The discussion chains consist of the paths in the graph whose starting node is a root and ending node is a leaf when we inverse the direction of the edges. In order to define a discussion chain, we consider $root(G)$ to be the set of vertices of the graph G which represent message objects that do not reply to another message. Moreover, $inReply(v)$ is the indegree set of vertices of the node v. A formal definition of a discussion chain follows:

Definition 3. We define a **discussion chain** in the graph $G = (V, E)$ as the subgraph

$$G_c = (V_c, E_c)$$

where
$V_c = \{v_i, v_{i-1}, v_{i-2}..., v_{i-x}\}$, $v_i \in root(G)$, $inReply(v_{i-x}) = \emptyset$, $v_i \neq v_{i-x}$,
$v_{i-k} \in inReply(v_{i-(k-1)}), \forall k, k \in [1, x]$ and
$E_c = (V_c)^2 \cap E$.

Similarly to the discussion threads, the discussion chains can also be queried by a specific message or user. The discussion chains where a message m appears are all the chains G_c of the graph G for which $\{\exists v \in V_c : v = (m, op_v, u_v, tm_v)\}$. Similarly, the chains where the user u has participated are the chains G_c of the graph G for which $\{\exists v \in V_c : v = (m_v, op_v, u, tm_v)\}$.

The distinction between a discussion thread and a discussion chain becomes apparent from Fig. 3 that shows a graph consisted of 2 discussion threads.

In Fig. 3, the first thread is consisted of 3 discussion chains:
$\{msgObj1, msgObj3, msgObj6\}$, $\{msgObj1, msgObj4\}$,
$\{msgObj1, msgObj2, msgObj5\}$. The second thread is consisted of 2 discussion chains: $\{msgObj10, msgObj11, msgObj13\}$, $\{msgObj10, msgObj12\}$.

The chains are important in a PROG graph. The longest discussion chain can point out the longest exchange of messages in a forum and it can be measured by the maximum number of edges that start from a leaf node and end up to a root node.

If we have more than one chain in the graph, then there is at least one node v that has received more than one reply. Additionally, if there exists a node $v \in V$ for which its reply has received another reply, then we assume that we have a generation of possible subdiscussions that start from v. Otherwise

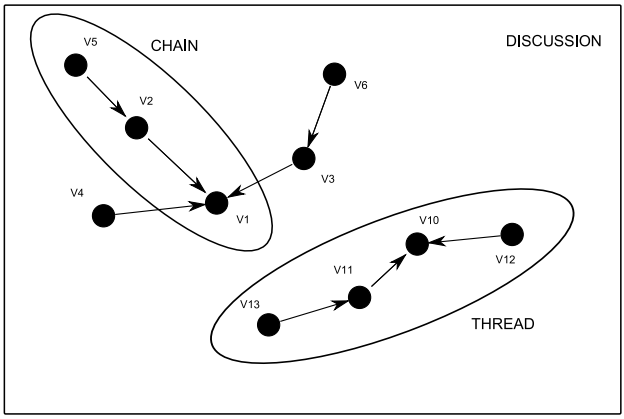


Fig. 3: Discussion threads and chains of a discussion

we consider to have only reactions and not subdiscussions starting from v. These arguments become clearer in Fig. 4.

For example, in Fig. 4 we can assume that the root of the graph (which, in this case, it is only one when we inverse the direction of the edges) has caused the generation of two different sub-discussions, one of which is initiated by the vertex in black. This black vertex, in turn, is dividing the discussion into two parts. The light grey vertex has caused four reactions that have not moved the discussion forward, so we cannot assume that there are four different arguments or sub-discussions that have been invoked.

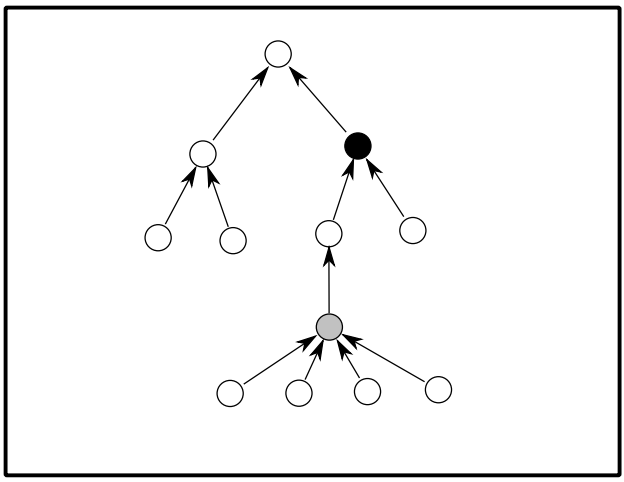


Fig. 4: A discussion thread in which the black vertex may have possibly caused sub-discussions while the light grey one has just caused reactions

4 Measures

After having defined the PROG model, we now define some measures that allow us to navigate inside an online discussion and analyze it efficiently.

4.1 Structure-oriented Measures

The PROG graph, by its definition, is not *complete* and it does not have the presence of *cycles* or *cliques* in it. Thus, theorems and measures that are defined in the theory of graphs regarding these concepts cannot be applied in our case. Nevertheless, there are some structural and elementary concepts of graph theory that are used in PROG graphs.

- **Identify the vertices of a discussion that initiate threads:** A root vertex is considered to be the vertex that initiates a discussion thread, even if this thread is consisted of only one vertex. The root messages of a graph $G = (V, E)$ are defined as:

$$root(G) = \{v \in V : outDegree(v) = 0\} \tag{2}$$

- **Identify the direct predecessors of a vertex:** The set of direct predecessors of a vertex is the set of reply nodes *inReply* towards a vertex v. According to the theory of graphs, it is defined as:

$$inReply(v) = \{v' \in V \mid (v', v) \in E\}. \tag{3}$$

  The number of the predecessors $inDegree(v) =\mid inReply(v) \mid$ shows how many reactions have been caused by the posting represented by the vertex v. This is a measure of the popularity of a message object and it can be an indicator for a classification of the postings from the most to the least popular. Popular postings point out the "heart" of a forum and their identification facilitates the mining of a forum by directing the analysis towards the most popular messages.

- **Distinguish the vertices which contain opinion inside a discussion chain:** Let us consider a discussion chain $G_c = (V_c, E_c)$ of the graph G. We can calculate the number of vertices that point out negative (n) opinions, objective (o) statements and positive (p) opinions respectively as:

$$vertCh(G_c, r) =\mid \{v \in V_c : op_v = r\} \mid \tag{4}$$

- **Find out the postings sent by a certain user:** Each forum participant may post many messages inside a forum. These messages are encapsulated in the message objects represented by the vertices of the graph. The message objects written by user u are given by:

$$msgs(u) = \{v \in V : v = (m_v, op_v, u, tm_v)\}. \quad (5)$$

4.2 Opinion-oriented Measures

In this Section, we concentrate on the measures that enable us to determine the flow of the opinion inside a discussion as well as the opinion status of the participants.

- **Identify the opinion status of a user onside a discussion:** A user may own more than one posting inside a discussion by replying, for example, to messages s/he has already received.

  Definition 4. The *opinion status* of a user is defined by the average opinion polarity s/he expresses inside the discussion.

  The opinion status of a user can be measured per discussion chain, per discussion thread and per discussion as a whole. In this way, we can observe users that keep a negative or positive position throughout the discussion or we can identify tendencies such as whether people tend to write more when they are unhappy or when they are satisfied with a certain situation. We define the average opinion expressed by a user u inside a discussion as: if $\mid msgs(u) \cap V_i \mid > 0$, then

$$avgOpFromUsr(G_i, u) = \frac{\sum_i op_{v_i}}{\mid msgs(u) \cap V_i \mid} \quad (6)$$

  where $v_i \in msgs(u)$ and $msgs(u)$ is given by the equation 5.
  The graph $G_i = (V_i, E_i)$ represents the subgraph of the discussion chain, the discussion thread or the graph of the whole discussion respectively according to what status we want to measure.

- **Identify the opinion of the reactions towards the postings of a certain user:** A posting written by a user may be replied to many times or it may be ignored. For the purpose of identifying the opinion status of other users towards a particular user u we define the average opinion expressed towards the user u (having received at least one answer) during the discussion as:

$$avgOpToUsr(G_i, u) = \frac{\sum_i op_{v_i'}}{\sum_j inDeg(v_j)} \tag{7}$$

where $v_j \in msgs(u) \cap V_i$, $v_i' \in inReply(v_j)$ and $inDeg(v_j) =\mid inReply(v_j) \cap V_i \mid$.
Again, the graph $G_i = (V_i, E_i)$ represents the subgraph of the discussion chain, the discussion thread or the graph of the whole discussion respectively according to what status we want to measure.

- **Identify the opinion of the reactions towards a certain posting:** The opinion polarity expressed towards a specific posting can be measured in many ways:

  1. *Number of replies towards the posting according to their opinion polarity.*
     A message object $v \in V$ may be replied to during a discussion through postings. These postings may contain objective information or they may include the sentiments of the author expressed by positive or negative opinions.
     We can always distinguish between the different postings according to their opinion polarity. The number of positive postings towards a message object $v \in V$ is defined by the number of reply vertices that contain a positive opinion. We describe the number of negative (n), objective (o) and positive (p) replies respectively as:

$$reply(v, r) =\mid \{v' \in inReply(v), op_{v'} = r\} \mid \tag{8}$$

  2. *The average opinion expressed towards the posting.*
     Measuring the average opinion received by a message object v can give us an indication of the reactions of the participants towards the specific posting. If, for example, the average opinion is 0, this means that either the reply postings contained objective information, or there is a balance between positive and negative opinions.
     We define the average opinion received by a message object $v \in V$ that has caused reactions as:

$$avgMsgOpinion(v) = \frac{\sum_i op_{v_i'}}{inDegree(v)}, \tag{9}$$

     where $v_i' \in inReply(v)$.
     The $inDegree(v)$ points out how many replies the posting represented by the vertex v has received.

3. *The variety in opinion polarity of the replies towards to the posting.* Having described the various vertices according to the opinion polarities included in their reply postings, allows us to define a measure regarding the *opinion information* held by a vertex. We use the entropy H for this purpose, and we define the amount of opinion information held by a vertex $v \in V$ (that has been replied to), as:

$$H(v) = -\sum_{r=n,o,p} \left(\frac{reply(v,r)}{inDegree(v)} \log \frac{reply(v,r)}{inDegree(v)}\right) \tag{10}$$

The opinion information measured by the entropy is used in general for measuring the diversity of opinions [10]. In our case it is an indication of the variety of opinions received by a vertex. If, for instance, a vertex has received reply postings that are all of the same opinion orientation, then the entropy will be 0. This may mean either that there is objective information or that there is unanimous opinion regarding the message expressed by the particular vertex.
The entropy is a measure that exploits the global (through the edges) together with the local information of the PROG graph. It is based on how the postings are linked to each other and on what opinion information they hold. A discussion analyst can distinguish the messages with high entropy among all others since these would be messages that have caused an intense discussion with various opinions.
Similarly to the *opinion information* measure per vertex $H(v)$, we can define the same measure per discussion chain. This measure facilitates the identification of the discussion chains that contain the maximum amount of opinion information.
The opinion information inside a discussion chain $G_c = (V_c, E_c)$ is defined as:

$$H(G_c) = -\sum_{r=n,o,p} \left(\frac{vertCh(G_c,r)}{\mid E_c \mid} \log \frac{vertCh(G_c,r)}{\mid E_c \mid}\right) \tag{11}$$

where n, o and p point out the negative, the objective and the positive vertices respectively and $vertCh$ is given by the equation 4.
The opinion information is an indication of the variety of opinions inside a discussion chain. Similarly we can define the opinion information inside a discussion thread.

4.3 Time-oriented Measures

The temporal dimension can facilitate the analysis of a discussion over a time period. The temporal information is mostly exploited in relation to other in-

formation such as the topic and the opinion, and hence it is used within other measures in other sections. In this Section we present only measures which are independent from the topic and opinion knowledge.

- **Identify the past of a posting beyond structure:** The temporal information permits to distinguish the messages that have been posted immediately before a specific posting. This cannot only be derived by the vertices which are counted in the *outDegree* of a vertex. The reason is that some users choose to send messages as a new posting without replying to an existing one, even though they may have been influenced by the existing postings. The case has been depicted in Fig. 2.
  We can define the ancestors of a vertex v as:

$$anc(v) = \{v' \in V : tm_{v'} < tm_v\} \tag{12}$$

4.4 Topic-oriented Measures

We can exploit the content of the different postings by using topic-identification algorithms [12].

Let the topic of a posting represented by a vertex v be denoted as $topic(v)$. Let us also consider a topic T.

Then, the postings that belong to this topic are given by:

$$msgs(T) = \{v \in V : topic(v) = T\}. \tag{13}$$

- **Find out the participation of a user in a certain topic:** For the purpose of identifying whether a user has had messages inside a topic as well as the proportion of them, we define the following measure:

$$userParticip(u, T) = \frac{\mid msgs(u) \cap msgs(T) \mid}{\mid msgs(u) \mid} \tag{14}$$

- **Identify the opinion evolution of a user in a topic:** A significant measure is that of the opinion evolution of a person in a particular topic. This measure becomes even more important if we know or have identified the profile of the user. If, for example, the user is an expert in the domain of the discussion, then his opinion has a higher value.
  This concept can be measured in various ways:

  1. *Average Opinion of a User in a Topic.* We can approximate this measure by observing the average opinion that a user has expressed in postings

that belong to a particular topic T, if $userParticip(u,T) > 0$:

$$opEvolution(u,T) = \frac{\sum_i op_{v_i}}{\mid msgs(u) \cap msgs(T) \mid} \tag{15}$$

where $v_i \in msgs(u) \cap msgs(T)$.

2. *Difference in opinion polarity of a User u between two timestamps of the same Topic T.*

$$opEvolution(u,T,tm_v,tm_{v'}) = \mid op_v - op_{v'} \mid \tag{16}$$

where $v, v' \in msgs(u) \cap msgs(T)$ and $v = (m_v, op_v, u_v, tm_v)$, $v' = (m_{v'}, op_{v'}, u_{v'}, tm_{v'})$.
The measure will give a 0 result if the user has not changed opinion between the two instances. It will be 1 if the user has moved from or to an objective posting and it will give the value of 2 if the user has gone from a positive to a negative opinion and vice versa.

- **Identify the opinion polarity expressed for a certain topic:** The opinion can also be calculated on average per topic without specifying a user. Then, the average opinion polarity in the discussion for a topic T is:

$$avgOpTopic(T) = \frac{\sum_i op_{v_i}}{\mid msgs(T) \mid} \tag{17}$$

where $v_i \in msgs(T)$.
In the same way, we can measure the average opinion polarity for a topic inside a discussion chain G_c or a discussion thread G_{thr}.

- **Identify how popular a topic is:** The popularity of a topic can be measured inside a discussion represented by the PROG graph $G = (V, E)$ as:

$$topicPop(G,T) = \frac{msgs(T)}{\mid V \mid}. \tag{18}$$

The popularity of a topic can be measured accordingly inside a discussion thread or even a discussion chain.

- **Identify the past of a posting by topic beyond structure:** Having the information of which topic a posting belongs to enables us to see the actual ancestors of a posting, not only by time (as in the time-oriented measures) but also by topic.

$$anc(v,T) = \{v' \in V : tm_{v'} < tm_v, topic(v') = topic(v)\} \quad (19)$$

where $v = (m_v, op_v, u_v, tm_v)$, and $v' = (m_{v'}, op_{v'}, u_{v'}, tm_{v'})$.
This measure can facilitate the identification of links between messages that are not captured by the structure of the discussion.

- **Identify the future of a posting by topic beyond structure:** Similarly to the ancestors, we can define the descendants of a posting by topic. This could be also seen as a measure of popularity for a vertex v, considering as "reactions", the vertices that have followed in time and they also belong to the same topic T as the specific vertex v. We define the descendants of a vertex v as:

$$descendants(v) = \{v' \in V : tm_{v'} > tm_v\} \quad (20)$$

where $v = (m_v, op_v, u_v, tm_v)$, and $v' = (m_{v'}, op_{v'}, u_{v'}, tm_{v'})$.

A summary of the presented measures that are based on the Post-Reply Opinion Graph is presented in Table 1.

5 Application

In this Section we show how the proposed model can be applied to a real forum and what information the measures we have defined provide us with.

We have taken a forum from the site of a French newspaper[1] that consists of 121 messages and 97 users. We have manually identified the opinion polarities since we have not found an available Opinion Mining tool that identifies opinion polarities in French text, and we have automatically created the PROG graph that is shown in Fig. 5. For the visualization of the graph we have used the JUNG library (http://jung.sourceforge.net). The vertices appear with an identification number calculated internally by our application developed for the purpose of visualizing and analyzing forums.

As we can see in Fig. 5, the PROG graph consists of some nodes that do not connect to the rest of the graph. These vertices are the ones in black and they represent message objects that they either do not reply to any other message or that they have not received any reply. These vertices are not very interesting for the discussion analysis since they have not played a crucial role for the development of the discussion. They are less probable to have an impact on the whole discussion or to contain interesting opinions. Visualizing the forum structure by a PROG graph allows us to concentrate on the discussion threads that consist of many vertrices or many discussion

[1] http://www.liberation.fr

Table 1: Measures for a discussion represented by a PROG graph G

Description	Formula
Root postings	root(G) $= \{v \in V : outDegree(v) = 0\}$
Popularity of a posting v	$inDegree(v) =\mid \{v' \in V : (v', v) \in E\} \mid$
Opinion status of a user u	$avgOpFromUsr(G_i, u) = \frac{\sum_i op_{v_i}}{\mid msgs(u) \cap V_i \mid}$
Opinion reactions towards a user u	$avgOpToUsr(G_i, u) = \frac{\sum_i op_{v'_i}}{\sum_j \mid inReply(v_j) \cap V_i \mid}$
Opinion reactions towards a posting v	$reply(v, r) =\mid \{v' \in inReply(v), op_{v'} = r\} \mid$ $avgMsgOpinion(v) = \frac{\sum_i op_{v'_i}}{inDegree(v)}$ $H(v) = -\sum_{r=n,o,p} (\frac{reply(v,r)}{inDegree(v)} \log \frac{reply(v,r)}{inDegree(v)})$
Opinion information of a discussion chain G_c	$H(G_c) = -\sum_{r=n,o,p} (\frac{vertCh(G_c,r)}{\mid E_c \mid} \log \frac{vertCh(G_c,r)}{\mid E_c \mid})$
Ancestors of a posting beyond structure	$anc(v) = \{v' \in V : tm_{v'} < tm_v\}$
Postings that belong to a topic T	$msgs(T) = \{v \in V : topic(v) = T\}$
Participation of a user u in a topic T	$userParticip(u, T) = \frac{\mid msgs(u) \cap msgs(T) \mid}{\mid msgs(u) \mid}$
Opinion evolution of a user u in a topic T	$opEvolution(u, T) = \frac{\sum_i op_{v_i}}{\mid msgs(u) \cap msgs(T) \mid}$ $opEvolution(u, T, tm_v, tm_{v'}) =\mid op_v - op_{v'} \mid$
Opinion expressed for a topic	$avgOpTopic(G_i, T) = \frac{\sum_i op_{v_i}}{\mid msgs(T) \cap V_i \mid}$
Topic Popularity	$topicPop(G_i, T) = \frac{msgs(T) \cap V_i}{\mid V_i \mid}$
Ancestors of a posting by topic	$anc(v, T) = \{v' \in V : tm_{v'} < tm_v, topic(v') = topic(v)\}$

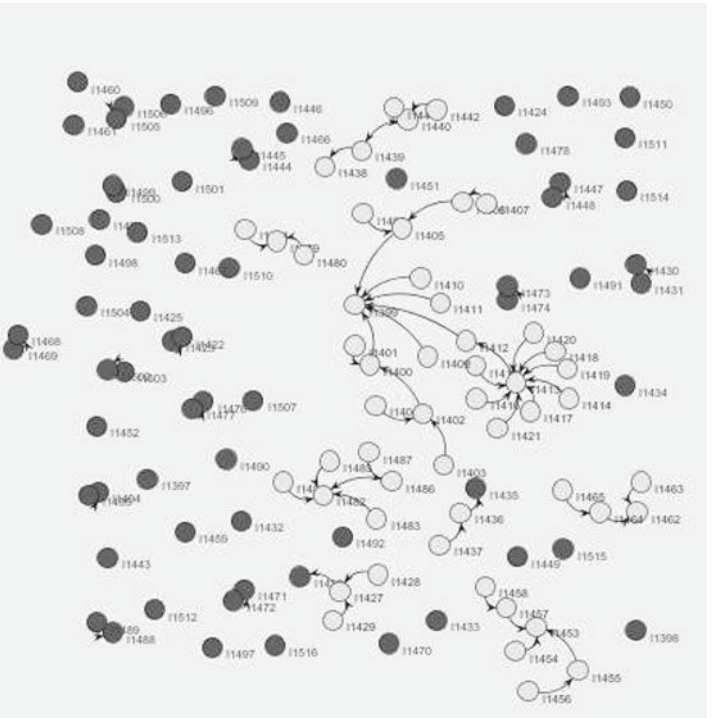


Fig. 5: PROG graph of a real web forum

chains. Such threads appear in the center of Fig. 5 and their vertices are colored in a light grey. The "black" vertices can be put on the side, for the time being, in an attempt to perform in this way a first step in the discussion mining.

The proposed model and the defined measures allow us to extract the following information:

- the postings that have initiated the existing discussion threads
- the most popular message m_p and which user u_p has written it
- what is the average opinion towards the message m_p
- how varied are the opinions expressed in the replies towards the m_p
- if the author of the m_p has written other messages and what on average is his/her general opinion status
- the average attitude of the rest of the authors towards the user u_p during the forum
- the discussion thread and the discussion chain that contain the message m_p or any other posting
- whether the particular discussion thread contains another popular message or not

- what is the message that has led to the most popular message m_p
- whether there are subdiscussions in the discussion thread that contain the m_p
- the opinion evolution of the topic to which the m_p belongs
- the real ancestors (beyond structure) from which the m_p or any other posting may have been influenced from.

This information cannot be given by the social network model of the forum, but it can be extracted from the Post-Reply Opinion Graph.

First of all, by looking at the content of $root(G)$ messages we can get a quick summary of the main aspects discussed in the particular forum. Let us, now, use the structure-oriented measures in order to "zoom" more into the discussion. By calculating the $inDegree$ of the vertices of the graph G, we can identify the most popular messages. In Table 2, we present them in descending order by the number of reactions they have received. We refer to them by the unique code given by our application and we provide also information regarding the average opinion of their replies and their variety given by the entropy.

Table 2: Most popular messages of the forum

v	$\mid inReply(v) \mid$	$avgMsgOpinion(v)$	$H(v)$
I1413	8	-0.375	0.39
I1399	6	-0.5	0.3
I1482	4	0	0
I1453	3	0	0.48

From Table 2, we can see that the message I1453 has the highest entropy of all. Indeed this is the message that has received replies with the highest variety of opinions. The message I1482 has 0 entropy which shows lack of opinion variety in the replies it has received. In combination with the value 0 of the $avgMsgOpinion$ we understand that all the replies are objective so they contain no opinion (otherwise the entropy would not be 0).

The extraction of the most popular messages as well as their opinion entropies and average opinions can help a user who has just entered the forum to get some information immediately without having to browse the whole forum. The user can read the popular messages to get an idea of what the participants are talking about. By the messages with the high entropy, the user can see the messages which have caused different opinion reactions.

The author u_v of each message m_v is known since the message and its author are both encapsulated in the concept "message object". The fact that we know the author of the most popular messages, allows us to look at the position and role of the particular authors in the social network model and derive some conclusions about them, such as how much they can influence the rest of the discussion, the communities in which they exist etc.

In Table 3 we show the results of the opinion measures oriented towards the authors of the most popular messages. Instead of giving the actual pseudo of each user, we name them "A", "B", "C", "D". From this table, we notice that the average opinion of user "A" is $avgOpFromUsr(u) = -1$. This shows that whenever this user wrote a message he expressed a negative position. At the same time the replies towards this user have had an average $avgOpToUsr(u) = -0.44$ which shows that the position of the rest of the users that replied to this one was negative as well. The user "D" has a balanced attitude (sometimes positive, sometimes negative).

Table 3: Measures applied to the users

Msg Object	User	$avgOpFromUsr(u)$	$avgOpToUsr(u)$
I1413	A	-1	-0.44
I1399	B	0	-0.5
I1482	C	0	0
I1453	D	0	-0.33

Regarding the discussion threads and chains, they are easily identified by the PROG model. We notice that both popular messages I1399 and I1413 belong to the same thread. This encourages us to give priority in analyzing the particular thread among all the rest, since its content should be more interesting having caused a discussion around it. Additionally, from our model we can distinguish the message that has led to the most popular message, just by following the respective edge that relates the specific message with its successor (predecessor in time). Finally the particular thread is divided into two subdiscussions.

In conclusion, the application of the PROG model to a forum results in the extraction of knowledge that cannot be provided by the social network model. More specifically, from a graph of PROG type we can extract the following information:

- The discussion chains and threads of the analyzed discussion.
- The popular postings which have caused many reactions.
- The opinion polarity presence in the discussion.
- The indirect links between the postings by using the temporal and the topic information.

These summarized points show the worth of the proposed model in the domain of discussion analysis.

6 Conclusion and Future Work

This article presents a theoretical work that consists in defining formally a Post-Reply Opinion Graph. The main novelty of our proposal is that we integrate into the model structure information of the discussion and the opinion content of the exchanged forum postings, information that is lost when we represent a forum by a social network model. We define measures that give information regarding the opinion flow and the general attitude of users and towards users throughout the whole forum. The application of the proposed model to real forums shows the additional information that can be extracted and the interest in combining the social network and the PROG models.

We believe that the future in Post-Reply Opinion Graphs is prosperous. Future work will pass from the theoretical to an experimental state by performing large-scale experiments with real forums. The information extracted by the PROG model can be used in many ways. We have experimented by using it in order to rank forum messages from the most to the least interesting. This is a combination of many criteria such as how many reactions a message causes, whether it receives reactions that contain opinions, whether these opinions have the same strength or not. Initial results are promising but more extensive experiments are needed.

One future objective is to exploit the time dimension in our model. This will permit monitoring how opinion changes over time. In this way, we could observe whether a product improves as the time passes, whether people become more satisfied with certain services, or even whether people are finally convinced after a long discussion in a forum.

Furthermore, an interesting future issue is to combine the social network and the PROG models for an improved discussion analysis. For example, we could extract the experts [15, 8] of the discussion domain through the social network representation, and we could, then, use this information in order to extract from the PROG model their attitude or the discussion chains in which they have participated.

The structure of the Post-Reply Opinion Graph allows the extraction of discussion threads and chains. This knowledge could be used in the future in order to give confidence to topic identification algorithms. For example, two messages that appear in the same discussion chain or thread have higher probability to belong to the same discussion topic. As a result, a topic-identification algorithm could give higher probability of belonging to the same topic to messages that do not only have similar content but they are also linked in the same chain or thread.

In the presented work, the relations between the messages are considered to be known (which message replies to what) but often these relations are not available. We have described a way on how topic and temporal information can aid in the identification of the real ancestors/descendants of a posting. In the future, we intend to work more with the automatic extraction of these

relations between postings and the population of the graph with appropriate links.

Additionally, our model captures currently cases where one message can reply to one and only one message. In some cases, though, one message may respond to more than one message. Future work needs to cater for changes in the model and the measures in order to capture this particularity.

References

1. R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining newsgroups using networks arising from social behavior. In *Proc. of the 12th International conference on World Wide Web*, 2003.
2. P. Carrington, J. Scott, and S. Wasserman. *Models and Methods in Social Network Analysis*. New York: Cambridge University Press, 2005.
3. X. Ding and B. Liu. The utility of linguistic rules in opinion mining. In *SIGIR-07*, 2007.
4. D. Fisher, M. Smith, and H. Welser. You are who you talk to: Detecting roles in usenet newsgroups. In *Proc. of the 39th Annual HICSS*. IEEE Computer Society, 2006.
5. A. Ghose, P. Ipeirotis, and A. Sundararajan. Opinion mining using econometrics: A case study on reputation systems. In *ACL*, 2007.
6. V. Hatzivassiloglou and K. Mckeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics*, pages 174–181, 1997.
7. M. Helander, R. Lawrence, and Y. Liu. Looking for great ideas: Analyzing the innovation jam. In *KDD*, 2007.
8. M. Hu, E.-P. Lim, A. Sun, H. Lauw, and B.-Q. Vuong. Measuring article quality in wikipedia: Models and evaluation. In *CIKM '07*. ACM, 2007.
9. M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD International conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.
10. K. Krippendorff. *Information Theory. Structural Models for Qualitative Data*. Sage Publications, 1975.
11. J. Scripps, P.-N. Tan, and A.-H. Esfahanian. Node roles and community structure in networks. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 26–35. ACM, 2007.
12. F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.
13. A. Stavrianou, J. Velcin, and J.-H. Chauchat. Definition and measures of an opinion model for mining forums. In *2009 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 09)*, 2009.
14. P. Turney and M. Littman. Measuring praise and criticism: inference of semantic orientation from association. *ACM TOIS*, 21(4):315–346, 2003.
15. J. Zhang, M. Ackerman, and L. Adamic. Expertise networks in online communities: Structure and algorithms. In *Proc. of the 16th International conference on World Wide Web*, pages 221–230, 2007.

Socio-contextual Network Mining for User Assistance in Web-based Knowledge Gathering Tasks

Balaji Rajendran and Iyakutti Kombiah

Abstract Web-based Knowledge Gathering (WKG) is a specialized and complex information seeking task carried out by many users on the web, for their various learning, and decision-making requirements. We construct a contextual semantic structure by observing the actions of the users involved in WKG task, in order to gain an understanding of their task and requirement. We also build a knowledge warehouse in the form of a master Semantic Link Network (SLN) that accommodates and assimilates all the contextual semantic structures. This master SLN, which is a socio-contextual network, is then mined to provide contextual inputs to the current users through their agents. We validated our approach through experiments and analyzed the benefits to the users in terms of resource explorations and the time saved. The results are positive enough to motivate us to implement in a larger scale.

1 Introduction

The availability of all kinds of information on the Web, along with powerful search engines enables the users to involve in Web-based Knowledge Gathering tasks. However these tasks differ from other typical information seeking tasks as they involve a process of learning to be accomplished by the user. We refer such tasks as Web-based Knowledge Gathering (WKG).

Balaji Rajendran,
Centre for Development of Advanced Computing,
68, Electronics City, Bangalore, India
e-mail: balajirajendran@gmail.com

Iyakutti Kombiah,
CSIR Emeritus Scientist, School of Physics,
Madurai Kamaraj University, Madurai, India
e-mail: iyakutti@gmail.com

WKG is a special kind of information seeking task that typically requires information foraging from multiple sources and correlating them cohesively in order to satisfy an information need. The WKG task is generally carried out by a user who is new to those particular domain concepts and would like to enhance his knowledge, leading to many explorations on the Web, during the task. Typically WKG tasks require research on a particular subject and the users involved in it aim to answer questions of the type 'How', and 'Why', instead of 'Where' and 'What'. The knowledge sought by the users would be of either analytical or embrained [1] in nature rather than encyclopedic [2] or operational knowledge.

The users of these WKG tasks intend to learn the concepts and their nittygritty's involved in respect to the given task. The users primarily like to avoid exploring unrelated resources and thereby deviating from the main objectives of their task. Also, the users would like to gain as much knowledge related to the task within a short period of time. The information seeking mechanisms and information foraging behavior [3] of the users would easily make the task complicated and hence require expert assistance. An idealistic approach to provide user assistance in WKG tasks would be to analyze and understand all the content available on the Web and then to gather the user's need explicitly from them and finally map it. As of today, this approach seems infeasible, because of the challenges in writing programs that can understand all kinds of contents and be conversant with all subject domains, and also the assumption that the users can exactly define what they are looking for. This approach however can be partially feasible, once the present day web gets converted to a full-fledged semantic web [4].

We present our approach based on the premise that users repeatedly involve in same or similar WKG tasks. We construct mechanisms that learn from the first user, and assist the later users who are involved in same or similar tasks and also continuously enrich the knowledge of the system with every subsequent user. So, a user involving in a WKG task that had been attempted by several other users would receive better user assistance. We associate agents with every user involved in a WKG task and these agents intend to assist their users by providing hints on the probable resources, and appropriate keywords to use in respect to their task. The main intention therefore is to reduce the cognitive load experienced by the users, in WKG tasks and also to save their time in exploring unrelated resources.

The rest of the chapter is organized as follows. The next section introduces some of the related work in the domain of user assistance. Section 3 explains our contextual approach in detail, including the main algorithm and socio-contextual mining. Section 4 validates our approach through experiments and implicit user analysis. Section 5 concludes the chapter.

2 Related Work

The means of assisting the users on the web had been studied for a long time since now. However, in the earlier days, the research was focused on assisting users to browse and navigate, such as Letizia [5] that proactively fetched links from the page currently being viewed by the user, and recommended those links that may be of interest to a particular user, by analyzing their activities of browsing, Web-watcher [6] that searched the web autonomously on the behalf of the user, and provided interactive assistance to the user, using machine learning techniques. More such systems, of agent-based user-assistance for browsing and searching can be found in [7]. The advent of sophisticated search engines like Google had solved this problem, with users using them to reach their favorite web pages.

However, with the vast explosion of information available on the Web and with its growing billions of users, the need for highly relevant results keeps growing. The information requirements of the users had also evolved, with the users now aiming to perform research or investigative tasks through Web, in order to gather knowledge. This has lead to the vigor in the research on assisting users in various Web-based information seeking tasks such as Web-based knowledge gathering or explorative learning tasks. The research is generally addressed from two stand points - One that focuses on understanding the user's information needs or intentions [8], and the other that aim for contextual information retrieval [9] from the Web documents [10].

While the former uses log analysis and other implicit techniques [11], the latter is modeled from the Information Retrieval perspective, and do not generally consider the information needs of the user. However the importance of context in retrieving documents became clearly established with research being focused on models for contextual mining [12].

A survey of the earlier information retrieval techniques on the web can be found in [13]. Later techniques such as collaborative filtering for recommending appropriate documents and Web pages have been studied in [14], which gives importance to those resources that are used most often by other user, gained popularity and also building explicit mechanisms for ratings of resources.

We believe that understanding the user's information need and connecting them to the contextually relevant content are mutually inclusive activities. In our proposed technique we implement implicit contextual mechanisms for understanding the user's requirements and match them with the existing similar tasks in the system to provide contextual inputs to the user. Recently, the need for assistance to users in Web information retrieval is being explored as a separate topic as Web Information Retrieval Support Systems [15].

3 A Socio-contextual Approach to Contextual User Assistance in WKG

3.1 Challenges in WKG

Web-based Knowledge Gathering is typically perceived as a complex task by any given user, as it involves exploration of several resources, browsing and choosing the appropriate resources for further analysis, learning of new terms and concepts on the fly, and using them to gather more information. This complexity is compounded by the cognitive loads imposed by the structure of the information retrieval processes on the Web. Also, as the users are typically new to the domain of the WKG task, they would struggle to find the right keywords and probably end up exploring many resources before finding the right resource.

Therefore user assistance in WKG needs to save the time and efforts of the users. This requires understanding of the user's information need. However an explicit means of obtaining the user requirement may not be all successful, because, in many cases, the users may themselves be unclear of their exact requirements. This demands other implicit methods, and we propose an implicit and contextual means of understanding the user requirement with respect to a WKG task.

3.2 Observing a WKG Task

Observation of a user's WKG task will give an understanding of what the user is intending to achieve. This is accomplished through agents - programs that will monitor the actions of the user, and take autonomous decisions. The first task of the agent is to monitor and gather the contextual information from the user involved in a WKG task. The agent primarily observes the keywords used by the user while searching, and the links explored by the user. These form the crux of the contextual information. The agents upon further observation organize them in the form of a Semantic Link Network (SLN) so that it could be later utilized for various decision-making purposes such as contextual retrieval, reasoning/inferring, and learning.

The agents also classify the actions of the users involved in WKG tasks into the following states [16]: Search - when the user enters a keyword in a search engine, Filter - when the user selects a resource for exploration from a set of results thrown by the search engine, and Gather - when the user uses a particular resource, to learn or gather information from it. We use the time spent factor on a particular resource to classify it under the gather state.

3.3 Semantic Link Network

Semantic Link Network (SLN) is a graphical model intended for organizing various web resources, for the purposes of learning and discovery by extending the hyperlink to a semantic link [17]. The SLN can be applied to any dynamic, decentralized and large-scale e-learning environments. SLN is a step towards the semantic web vision [4], and its fundamental concepts are described in the seminal work of [18].

We use SLN for modeling and representing the contextual information gathered through the actions of the users involved in WKG task. The SLN could then be mapped to form the outline of the cognitive structures that depict the knowledge required for understanding a concept. The SLN is flexible enough to model the assimilation and accommodation processes [19] of the cognitive structures that are essential for learning new concepts, and we simulate those processes to organize the contextual information during the construction of SLN. An agent acting on behalf of a user can thereby learn and infer from the SLN and provide contextual assistance.

3.4 Types of SLN

We propose to construct two kinds of SLN - singular SLN and master or reference SLN. Both the SLN would be composed of nodes, where each node would represent a Web resource and anchored or linked to a keyword. The singular SLN is constructed each time when a WKG task is undertaken by a user in the system. The agent corresponding to that user constructs the singular SLN as it observes the actions of the user - the keywords used, links explored, time spent on a resource, during their WKG task. The singular SLN therefore represents the contextual information gathered by the agent during a user's WKG task in an organized manner. The master or reference SLN is an aggregation of several singular SLN constructed by various agents for various tasks of their users. This is constructed by a master agent by receiving and integrating singular SLN from many agents. The assimilation and accommodation processes are implemented during the construction of master SLN by the master agent. Therefore the master SLN is the warehouse of knowledge, from which the contextual user assistance is provided. The user's agent during their construction of singular SLN make comparison with the master SLN in order to elicit more information about the needs of the user and to direct the user to appropriate resources, based on their current knowledge level. This master SLN plays an important role in the entire process of contextual assistance to the users, as it consists of several integrated and refined singular contextual structures from which the appropriate information has to be elicited.

3.5 Construction of SLN

First let us start with the construction of singular Semantic Link Network by a user agent. In a WKG task, the keyword used is not only important for the user, but also for the agent as it is the primary contextual information, based on which the various resources explored by the user can be connected to. Also, the comparison with the master SLN will help the agent to determine the kind of WKG task that the user is involved in, and would go on to reveal the subject domain and sub-domains, related with the task. We further leverage the keywords used by the user to understand the complexity of the subject domain and the user's current knowledge level in that domain.

The second parameter is the resource links explored and used by the user. The resources are created as nodes and anchored or linked to the keywords that were used by the user for reaching them. We do not analyze the content of the resources; instead we only track the URL of the resource and their connectivity with a keyword. The link between a node and a keyword also carries a weight factor that quantifies the semantic relationship between them. A schematic view of the SLN is given in Figure 1.

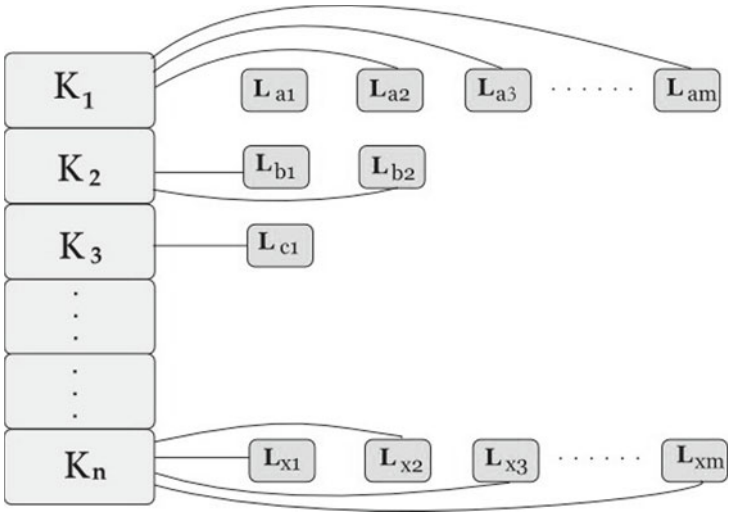


Fig. 1: Schematic View of a Semantic Link Network

We also track other ancillary parameters such as the time spent by the user, in a particular resource, which serves as an indication of relevancy and usefulness of that resource to that particular WKG task. The number of times a particular resource gets utilized, is given as the weight of the link between the resource and the anchored keyword, which gains value in case of the reference SLN.

A resource node in a SLN therefore constitutes all the above parameters. For instance a node RL_i is given as $K_i, A_i, W_i, TS_i, TW_i$ where K_i represents the key-word used by the user to reach this resource; Ai represents the URL address of the resource; Wi represents the weight indicating the weight of the resource link, based on the user's repeated visits; TS_i represents the timestamp of the last visit by the user; TW_i represents the total time spent on the resource by a user. We maintain an index of keywords to which resources are anchored. The following algorithm illustrates the process involved in inserting a node in the construction of a SLN. The same algorithm is used for integrating the singular SLN with the master SLN, by following a node-by-node insertion.

Algorithm 2 Inserting a node in SLN

Input: RL_i is the node to be inserted;
Output: Updated SLN

```
1: void insertNode (Node RL_i)
2: Begin
3: if K_i exists already in the keyword index table then
4:    Scan for nodes attached to K_i containing the same address as A_i
5:    if found then
6:          //Node with same keyword and same address
7:          Update W, TS, and TW values of that node and exit.
8:    else
9:       //Node with a same keyword, but with a new address
10:      Create a link from K_i to the node RL_j
11:   end if
12: else
13:   Check if there exists a node RL_k in the SLN such that
14:   RL_k → A_k = RL_i -> A_i
15:   if found then
16:      //Node with same address but with a new keyword
17:      Insert K_i into the keyword index table
18:      Create a link from K_i the node RL_k
19:      Update W, TS, and TW values of that node and exit.
20:   else
21:      //A new node - with a new keyword and a new address
22:      Insert K_i into the keyword index table
23:      Create a link from K_i the node RL_i
24:   end if
25: end if
26: End
```

3.6 Restructuring Master SLN

The master SLN is not a plain integration of all singular SLN, but a structured integration such that it becomes a well-organized warehouse of knowledge. This requires consolidation of information in the master SLN which is achieved through keyword and address reorganization.

Keyword Reorganization: During the integration of singular SLN with the master SLN, the master SLN does a reorganization of the keywords. This is because the users use many variants of a same keyword, resulting in the duplication of a string of resource nodes anchored across multiple keywords. In such cases, the least used keyword variant will be removed from the master SLN, by reanchoring its resources to the predominantly used keyword. Most importantly this reorganization helps in strengthening the master SLN by identifying similar and different concepts.

So, the keywords in the index table of a well populated and well organized master SLN would indicate distinct concepts. This process also ensures quick retrieval of the required relevant structures. Also, the nodes anchored to such well organized keywords would indicate the cluster of sub-concepts under them.

Address Reorganization: The presence of information in multiple pages of the same domain can result in generation of too many resource nodes. This dilutes the overall importance of the resource. In such cases the nodes are aggregated by cutting the right-most portions of the URLs to have the same address and thereby refer to the existing node.

The integration of singular SLN with the master SLN, followed by the keyword and address reorganization completes the simulation of the assimilation and accommodation processes, as it happens with the cognitive structures. This process ensures Self-learning by the master SLN and also prepares it for contextual mining of information. This refined master SLN forms the socio-contextual network that is ready for mining inputs to be provided to users executing similar tasks.

3.7 Socio-contextual Mining of Master SLN for Discovering Similar Structures

The comparison of singular SLN with the master SLN and inferring of information from it forms the crucial aspect in this entire process of providing contextual user assistance in WKG tasks. The user agents during the construction of their singular SLN seek inputs from the master agent by presenting the node structure. The task of the master SLN then is to mine the appropriate patterns from its rich structures and present the relevant result segments to the user agents. The user agent is allowed to seek for inputs only

after its singular SLN is populated with atleast one new keyword along with two resources, in order to have meaningful inputs from the master SLN.

The master agent uses the contextual parameters of the presented structure to determine the appropriate segment structures within it. The master agent extracts the keywords and subjects them to keyword reorganization, after which it scans for the appropriate matches in the index table. If a match is found, the corresponding keywords from the master SLN and their associated resources would be selected and presented to the user agent. If no match could be found, then based on the similarity of the resources, the master SLN would detect the appropriate keywords and the relevant resources, and present to the user agent.

The user agent uses the inputs obtained to find new leads in the WKG task in the form of resources and keywords that has not been so far explored by its user and presents it to them. However, the agent does not present all the new resources and keywords, but makes an estimate of the level or stage of the user in the WKG task, based on the singular SLN and the inputs obtained from master agent. If there are lots of resources, the user agent at the most presents those resources and keywords that would correspond to the next two levels in the WKG task. In other words, the user agent would not suggest more than two keywords at a time. The number of resources associated with those keywords would also be selectively presented. This is achieved by applying socio filters such as the last accessed time of a resource, and usefulness of a resource to that category of user, to refine its selections and create a ranked list of resources to be presented to its user. The user agent therefore is able to provide personalized and contextual assistance by suggesting the appropriate resources and keywords relevant to their user's current WKG task.

There would be cases where appropriate structure in master SLN could not be found, and the user could not be assisted. In such cases, the singular SLN will be consumed fully by the master SLN and would be provided as inputs to other users who later involve in similar tasks and their contextual task structures would be used for refining and enriching the existing structure.

4 Experiments and Evaluation

We developed an environment for WKG tasks that had interfaces to popular search engines, recorded the keywords and resources used, along with the time spent on a resource, and also had features for creating and saving notes on a particular topic. Also, the users can save their work, and can get back to the same point, at a later point of time. This feature is helpful in case of a medium to complex WKG task as the user would typically carry out such task in multiple sessions. In the preparatory phase, we constructed 10 WKG tasks spanning three diverse domains - Computer Network Security,

Architectural Engineering, and Cluster setup for research in computational physics. The last one was designed to have a more practical and hands-on task. We identified tasks that could be rated from easy to medium and presented it to a group of 10 candidates who were novices to that particular domain. The user agents observed the actions of the user and constructed the singular SLN by using the keywords used, and the resources explored by the user, and communicated it to the master agent. The process of retrieving similar structure, by the master agent, when a user agent communicates its singular SLN was switched off during this phase. Hence the users were not provided any assistance by the user agents that passively observed and collected their data and sent them to the master agent for constructing the master SLN. After the tasks were completed, the self-restructuring process of master SLN was activated and upon examination of the master SLN we found 21 keywords and 88 resource nodes.

4.1 Experiment

We designed 20 different WKG tasks spanning the three above mentioned domains, of which six were exactly the same tasks that were used during the preparatory phase, and nine of them required additional knowledge on top of the structure already present in the master SLN, while the remaining five tasks were new subtopics of the above domains. The tasks from the domain of Network Security (NS) required creating research reports on a particular technique, say Anomaly detection, and also included writing short program scripts, and tasks from the domain of Architectural Engineering (AE), required the candidates to understand a concept such as deep foundations, and included preparing documents and artifacts, while tasks from Cluster setup (CS) required users to derive specifications and configure and install software such as understanding ATLAS package and determining the pre-requisites for installing it, including the actual installation of it . The tasks were categorized at three inherent complexity levels [20] - Simple, Medium and Difficult. A difficult task is defined as requiring understanding of more than 3 basic concepts of the given domain.

A new group of 10 candidate users were selected and were given these tasks to be carried out. Each candidate was given two WKG tasks from different domains. The candidates were allowed to carry out the tasks in multiple sessions, but were given a deadline, before which they should have completed the tasks. The tasks were created out through the specialized environment developed for WKG tasks, and the actions of the users were recorded. The actions of the user agents such as their queries to master agent and their inputs to the user were also recorded.

4.2 Analysis

After the completion of all the tasks, by all the users, we evaluated the socio-contextual mining of patterns by the master agent from its master SLN to give inputs to the user agents. Table 1 gives the total number of requests made by user agents for a task of a particular domain and particular complexity, and the responses in terms of the number of nodes given by the master agent for each such request. It can be observed that the number of user agent requests is less for a simple task, and more for a difficult task. However it should be noted that this mainly depends on the users - as the users venture to gather more knowledge, more the requests would be made to the master agent.

The simple tasks are the one's where there would be enough information available with the master SLN, and the extraction of appropriate nodes involves the effective use of socio-contextual mining. The master agent typically tries to extract nodes from strongly related concepts also along with the matching appropriate concept. However, it refrains itself from selecting nodes from loosely-related concepts, and even less frequently used nodes within the same concept. This is to have better quality responses and as the user progresses, the user agent will be returning with new requests.

Table 1: User Assistance (UA) requests and responses from Master Agent

Task Domain	Task Complexity	No. of WKG Tasks	No. of UA requests received	No of nodes in each response	Average time spent for completion of the task in minutes
NS	Simple	1	2	(3, 3)	24
NS	Medium	3	5	(4, 3, 3, 2, 2)	57
NS	Difficult	2	7	(3, 3, 2, 2, 1, 0, 0)	108
AE	Simple	3	3	(4, 3, 4)	22
AE	Medium	2	9	(0, 3, 2, 1, 0, 1, 1, 0)	49
AE	Difficult	0	0	-	
CS	Simple	2	4	(2, 1, 2, 1)	27
CS	Medium	4	10	(3, 3, 2, 2, 1, 0, 1, 1, 2, 0)	66
CS	Difficult	3	11	(4, 4, 3, 2, 2, 1, 1, 1, 0, 0, 1)	142

The medium tasks were primarily designed to extend the existing knowledge of the master SLN, by introducing new additional concepts on top of the simple concepts. We found that the master SLN gained more knowledge in these medium tasks in terms of the number of nodes and concepts / keywords getting added to its existing structure. The master SLN now consisted of 36 keywords and 118 resources.

The difficult tasks involved learning and understanding of certain basic concepts also, some of which were covered by the simple WKG tasks. So, the users got many inputs in the initial stages of their difficult tasks, than in the later stages. Also, as the users' progress in their difficult task, they tend to move towards a more specialized concept that may not be covered

in master SLN so far, thereby having less number of responses. From our experiments, we found that the user assistance by the agents had saved time and the number of resources explored by the user. A comparison with the six of the WKG tasks that were used in the preparatory phase and repeated in the experimental phase is given in table 2. The results clearly indicate lesser number of resources used and significant time saved.

Table 2: Comparison of WKG Tasks with and without User Assistance

Task No	Preparatory Phase		Experimental Phase	
	Time Taken for Completion (Minutes)	No. of Keywords and Resources used	Avg. Time taken for completion (Minutes)	Avg. No. of Keywords and Resources Used
T1	36	6, 13	22	3, 8
T2	39	7, 11	24	3, 7
T3	89	13, 19	66	6, 14
T4	94	11, 14	71	7, 12
T5	204	18, 34	142	11, 18
T6	166	14, 32	108	9, 16

5 Conclusion

User assistance in a knowledge-intensive job is very challenging for any computing system and in cases of WKG tasks an expert's knowledge is typically required to assist the users and obtaining an expert's knowledge in every possible domain is highly infeasible. Hence we attempt to find a trade-off between no-assistance and expert assistance, by using socio-contextual mechanisms. Our approach is on the premise that a WKG task is repeatedly carried out by many users in slightly variant forms, and we tap into the power of social computing to provide user assistance. We aim to understand the user's information requirements and map it with the contextual knowledge gathered from several other users and assist the user in question. The experiments spanning multiple domains validated our approach by saving the user's time and linking them to the related resources for their tasks.

Acknowledgements

We express our earnest thanks to the anonymous reviewers, for their helpful comments.

References

1. Collins, H.: The structure of knowledge, Social Research, vol. 60, no. 1, pp. 95-116. (1993)
2. Fujii, A.: Organizing encyclopedic knowledge based on the web and its application to question answering. In: 39th Annual Meeting on Association for Computational Linguistics, pp.196-203. (2001)
3. Pirolli, P., Card, S. K.: Information Foraging, Psychological Review, vol. 106, no. 4, pp: 643-675 (1999).
4. Berners-Lee, T., Hendler, J., Lassila, O.: Semantic Web, Scientific American, 284(5) pp. 34-43, (2001)
5. Lieberman, H.: Autonomous Interface Agents. In: ACM Conference on Computers and Human Interaction (CHI-97), (1997)
6. Joachims, T., Freitag, D., Mitchell, T.: WebWatcher: A Tour Guide for the World Wide Web. In: International Joint Conference on Artificial Intelligence (IJCAI-97), Nagoya, Japan (1997)
7. Huhns, M.: Readings in Software Agents. Morgan Kauffman (1997)
8. Rose, D. E., Levinson, D.: Understanding user goals in web search. In: 13th international Conference on World Wide Web, WWW '04, ACM, New York, NY, pp.13-19, (2004)
9. Ingwersen, P., Järvelin, K.: The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series). Springer-Verlag, New York (2005)
10. Lawrence, S.: Context in Web Search. In: IEEE Data Engineering Bulletin, vol. 23, no. 3, pp.25-32 (2000)
11. Claypool, M., Brown, D., Le, P., Waseda, M.: Inferring User Interest. In: IEEE Internet Computing, November-December, pp.32-39, (2001).
12. Qiaozhu, M., ChengXiang, Z.: A Mixture Model for Contextual Text Mining. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2006) pp.649-655, (2006)
13. Kobayashi, M., Takeda, K.: Information Retrieval on the Web. In: ACM Computing Surveys, vol. 32, no. 2, pp.144-173, June 2000 (2000).
14. Hill, W., Stead, L., Rosenstein, M., Furnas, G.: Recommending and Evaluating Choices in a Virtual Community of Use. In: ACM Conference on Computers and Human Inte-raction (CHI-95), (1995)
15. Hoeber, O.: Web information retrieval support systems: The future of web search. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WIIAT'08), vol. 3, pp. 29-32. IEEE Press, (2008)
16. Rajendran, B.: Socio-Contextual filters for Discovering Similar Knowledge-Gathering Tasks in Generic Information Systems. In: Yang, C.C., Carley, K., Mao, W., Zhan, J., Chen, H., Chau, M., Chang, K., Lang, S., Chen, P., Hsieh, R., Zeng, D., Wang, F., (eds.) ISI 2008 SOCO Workshop. LNCS, vol. 5075, pp.384-389, Springer, Heidelberg (2008)
17. Zhuge, H.: Communities and Emerging Semantics in Semantic Link Network: Discovery and Learning, IEEE Transactions on Knowledge and Data Engineering, IEEE computer Society Digital Library, (2009)
18. Zhuge, H.: The Knowledge Grid. World Scientific Publishing Co., Singapore, (2004)
19. Piaget, J., Brown, T.: The equilibration of cognitive structures: The central problem of intellectual development, University of Chicago Press, (1985)
20. Rajendran, B., Kombiah, I.: Cognitive agents for understanding the complexities involved in web-based knowledge-gathering tasks. In: IEEE International Conference on System, Man and Cybernetics, pp: 1310-1315, (2009)

Integrating Entropy and Closed Frequent Pattern Mining for Social Network Modelling and Analysis

Muhaimenul Adnan, Reda Alhajj, and Jon Rokne

Abstract The recent increase in the explicitly available social networks has attracted the attention of the research community to investigate how it would be possible to benefit from such a powerful model in producing effective solutions for problems in other domains where the social network is implicit; we argue that social networks do exist around us but the key issue is how to realize and analyze them. This chapter presents a novel approach for constructing a social network model by an integrated framework that first preparing the data to be analyzed and then applies entropy and frequent closed patterns mining for network construction. For a given problem, we first prepare the data by identifying items and transactions, which are the basic ingredients for frequent closed patterns mining. Items are main objects in the problem and a transaction is a set of items that could exist together at one time (e.g., items purchased in one visit to the supermarket). Transactions could be analyzed to discover frequent closed patterns using any of the well-known techniques. Frequent closed patterns have the advantage that they successfully grab the inherent information content of the dataset and is applicable to a broader set of domains. Entropies of the frequent closed patterns are used to keep the dimensionality of the feature vectors to a reasonable size; it is a kind of feature reduction process. Finally, we analyze the dynamic behavior of the constructed social network. Experiments were conducted on a synthetic dataset and on the Enron corpus email dataset. The results presented in the chapter show that social networks extracted from

Muhaimenul Adnan
Department of Computer Science, University of Calgary, Calgary, Alberta, Canada

Reda Alhajj
Department of Computer Science, University of Calgary, Calgary, Alberta, Canada
Department of Computer Science, Global University, Beirut, Lebanon
Department of Information Technology, Hellenic American University, Athens, Greece

Jon Rokne
Department of Computer Science, University of Calgary, Calgary, Alberta, Canada

a feature set as frequent closed patterns successfully carry the community structure information. Moreover, for the Enron email dataset, we present an analysis to dynamically indicate the deviations from each user's individual and community profile. These indications of deviations can be very useful to identify unusual events.

1 Introduction

With the advancement in information technology such as the development of storage media, the improvement of the computing power of the commodity CPUs, and the availability of internet and broadband telecommunication; business organizations are now storing more and more data in electronic format in an attempt to add value to the online services provided to the customers. But keeping data in the raw format provide no or little decision making power to the managing body of the organizations. Hence, several knowledge discovery and lately data mining techniques (such as classification, clustering, association rule mining, and time series prediction) have been developed to facilitate the decision making of the organizations. These data mining techniques help the organizations to extract the information or knowledge buried in the data which is the key to the success of today's organizations. Most of the data mining techniques concentrate on finding information from data, or predicting future outcomes, but they provide little or no knowledge on how users of these data are connected to each other. We argue that social network is a powerful model for filling this gap. In other words, connecting people by considering their related information, connecting different information sources, and connecting people to information to which other people are connected are some of the important cases that may provide useful decision making power to the organizations, yet received little attention. In other words, an essential key factor to achieve success in a competitive environment is to avoid handling the information and people in isolation. In a social network settings information could be linked to other information and even to people. Actually, it is important to realize that people and the information are tightly coupled forming a social network which could be analyzed for interesting discoveries.

A social network represents relationships or ties among actors which depending on the problem domain could be individuals, pieces of information, genes, etc; actors are they key entities/items in the problem to be investigated. Formally, a social network can be described as a graph $G = (V, E)$, where $V = \{v_1, v_2, v_3, \ldots, v_n\}$ is the set of vertices representing individuals or actors and E is the set of edges representing relationships between vertices or individuals. In early days, the social networking concept was limited to analyze links between people in small communities like employees in an

organization, animals in a farm, actors in plays, etc; such social networks were mostly analyzed manually using classical graph theory, linear algebra and statistics. However, a large number of automated tools have been recently developed since the interest in the applications of social networks has received considerable attention in the research community with the currently emerging electronic forums and online friendship networks like facebook (www.facebook.com), orkut (www.orkut.com), myspace (www.myspace.com) etc. Consequently, it has been discovered that many complex and real world system from diverse fields can be modelled as social networks [25]. In fact, almost every aspect of the daily life may be modelled as a social network. This could be realized by deciding on actors and the links between them; actors may be homogeneous or heterogeneous. A homogeneous set of actors lead to a one-mode social network. On the other hand, a heterogeneous set of actors consisting of n different groups of actors lead to a n-mode social network; in practice the most common value for n is 2. In other words, we argue that it is possible to model many real world and research problems as social network; the main challenge would be the mapping of the problem domain onto a social network by deciding on: (1) the different categories of actors; one and two types are the most common in practice; (2) actual actors in each category; and (3) the links between the actors within the same category and across categories. For instance, a one-mode network may be constructed to study the relationship between students in a given course; a two-mode network may be used to model the relationship between software developers and the software projects.

A social network is mostly represented by adjacency matrix and folding is a common technique for deriving one-mode network from two-mode network; folding is realized by multiplying a given adjacency matrix and its transpose. Regardless of whether one-mode or two-mode, identifying communities of actors is an interesting problem in social networks [19, 12, 22, 8, 26, 21, 7]; the target is to find community structures from social networks. Knowledge of communities can help us in various applications such as product marketing, identifying web communities to make better web structure design, identifying organizational dynamics, etc.

In this chapter, we argue that modelling social networks from data sources is another important aspect of social networking research beside the community structure mining. In fact model construction is a vital step because everything else is based on the correctness and completeness of the model. Hence, it is necessary to invest more time and effort in the construction of the social network. Different social networks can be constructed from the same data sources based on different modelling scheme of the data, i.e., it is possible to produce different social networks from the same data by analyzing the data from different perspectives. For instance a corpus of documents may be the source to different social networks, including a social network of the documents based on the occurrences of words or phrases in the documents, a social network of the words/phrases based on their frequencies in the individ-

ual documents and across the documents, etc. But care must be taken so that when the social networks are extracted from the data sources they represent the underlying concepts accurately in a way that satisfies the purpose of the analysis. How a user of the data is connected to the information represented by the data should be the main focus. This chapter addresses these issues. We argue that the data to be analyzed in order to construct the social network can be transformed into a new space using data mining techniques, namely by mining frequent closed patterns [20]. Moreover, we can select only few frequent closed patterns (thereby decreasing the dimensionality) based on an entropy criterion for the modelling purpose; and this allows us to accurately discover useful communities from the analyzed dataset.

The frequent patterns mining model is general enough to be successfully applied to different domains. However, the key issue is deciding on the items and transactions for the particular problem to be solved. These are some examples of items and transactions from different domains: (1) items are software developers in a software company and a transaction is the set of software developers involved in the same project; (2) items are drugs and a transaction is the set of drugs that are effective for curing a certain disease; (3) items are genes and a transaction is the set of genes that act as markers for a given disease; (4) items are words or phrases in documents and a transaction is a document; (5) items are web pages of a particular website and a transaction is a set of pages visited during a user session. After the data is prepared, the closed frequent patterns are identified and then filtered using entropy.

To depict the usefulness of the proposed approach, we have conducted two sets of experiments. The first set of experiments is based on a synthetic dataset; and the other set of experiments analyzes the benchmark Enron corpus email dataset [6]. The results presented in this chapter show that the social networks extracted from a feature set as frequent closed patterns successfully carry the community structure information. Moreover, like the work of Wan *et. al.* [28], we conduct further analysis for the Enron email dataset to dynamically catch the email profile deviations from each user's individual and cluster profile. Scatter plots of these deviations as reported later in Section 6 can help us in identify unusual events.

The rest of the chapter is organized as follows. The related works are described in section 2. Section 3 presents the problem specification. The proposed framework is described in section 5. Section 6 provides the experimental results and the chapter is concluded in section 7.

2 Related Works

We know that a social networks is usually represented as a graph or network, sometimes called sociogram. Nodes in the graph represent the actors which are the entities or items to be studied. Each edge in the graph or network presents a relationship; and the strength of a relationship is depicted by the weight associated with the corresponding edge. Moreover, each edge can be directed or undirected representing an asymmetric or symmetric relationship. How the social network of relations is constructed differs from application to application. However, data mining and machine learning techniques can provide significant insight on the modelling of social networks based on data properties. For instance, in a social network of webpages, the relationship between actors (webpages) can be established by Google webpage frequency [17, 24]; the links between people in chat rooms can be deduced by using text-based segmentation of chat room conservations [13], or by using the influence diffusion model [16] where the relationship is based on frequency of terms propagated between two individuals. But most of the modelling methods are application specific and lack generalized guidelines for modelling the social network. In this chapter, we present a generalized framework for constructing a social network model using a well known data mining technique, namely frequent pattern mining [1], which is suitable for many application domains.

Once the social network model is created, the next task is to identify the interesting groups of communities from the network. There has been a significant amount of research literature that focuses on finding groups or communities from social networks. Based on the basic concepts of how the grouping is performed, they can be categorized into tow main branches: (1) partitioning based or graph theoretic methods, and (2) hierarchical methods. Some of the notable graph theoretic methods to identify communities from social networks include the approaches described in [19, 12, 14]. Some of them operate by repetitively bisecting the graphs to find the required number of communities or clusters; others use the concept of random walk to grow the communities from single node communities to larger ones such that the mean-squared distances between nodes of a community are minimized. One of the prominent graph theoretic methods is the K-means algorithm [14] which starts with randomly chosen initial set of k-partitions. It then calculates the centroids of all the partitions. Then, a new set of k partitions are constructed based on these new centroids. These two steps are iteratively repeatedly until the centroids converge, i.e., none of the data point will move to another cluster.

Hierarchical approaches [22, 8, 26, 21, 7] construct a hierarchy of clusters either using agglomerative or divisive techniques. Agglomerative methods start with an empty network and assign edges to networks based on some similarity measure. On the other hand, divisive methods repeatedly remove edges from networks. Girvan and Newman [8], presented a divisive hierarchi-

cal approach that works on the idea of edge betweenness. Edge betweenness is a measure, which presents the number of shortest paths(between vertices) that the edge takes part in. If betweenness for an edge is high, it is assumed that the edge is a connector of two loosely connected groups. Girvan and Newman [8] repeatedly remove edges with highest betweenness from a network until no edge remains. It has been observed, by the researchers that hierarchical approaches produce better quality grouping that is most suitable for social network analysis.

The idea of using frequent patterns for cluster analysis is not new. The work described in [3] proposed two text clustering methods based on frequent itemsets that appear in text documents. The FTC algorithm is used to find flat clustering and the HFTC algorithm is used to find the hierarchical clustering of the text documents. The authors argue that by using frequent patterns to describe the clusters, one can reduce the dimensionality of the document vector space by a significant amount which is otherwise large; this will have direct positive impact on the final outcome because dimensionality reduction is one of the major concerns to be tackled before the clustering process is applied on the data. Their idea is to find a clustering description as a subset of frequent itemsets such that the description covers the whole database. Members of the clustering description are chosen sequentially using a greedy method based on some entropy overlap score. In the proposed social network extraction framework presented in the literature, similar entropy measure was chosen to rank the significant features presented in the social network nodes. In the work described in [29], the authors used closed termsets to cluster text documents rather than using frequent itemsets. By using closed interesting patterns, they are able to reduce the dimensionality even more; and hence they are able to improve clustering quality.

We decided to consider the Enron email dataset as the real world case for demonstrating the effectiveness of the approach presented in this chapter because since the publication of the Enron email dataset [6], it has been the focus of the research efforts by several data mining research groups. For instance, Carenini *et. al.* [4] analyzed the Enron email data to show the effectiveness of their email conversation summarization method CWS, while Cselle *et. al.* analyzed the Enron data to detect topics from emails in order to better organize the emails. To discover the hidden organizational structure and the influential members in an organization, Shetty *et. al.* [23] proposed a graph entropy based method to identify a node's importance in a graph or social network. There has been many other research efforts described in the literature that analyze the other aspects of the Enron email data such as unusual email detection [11] based on deception theory, contextual search [18], unusual event detection [28], etc. In the work described in [28], the authors proposed a method where they can identify unusual events based on a user's individual and cluster mahalanobis distance. For our proposed method described in this chapter, we used a similar technique; but for our approach, we have used frequent closed patterns as features instead of using features such

as number of email recipients. Our approach is more robust because a closed frequent pattern combines a number of characteristics rather than restricting the actors and links to represent single individual characteristics from the domain.

3 Problem Statement

Given a set of data entities, it is required to construct a social network by considering the characteristics of the whole dataset. Formally, assume we have a set E of n entities $\{e_1, e_2, \ldots, e_n\}$. Each entity e_j is associated with a dataset D_j that represents the information regarding the entity e_j. Here, the dataset $D = \cup_j D_j, \forall j \in \{1..n\}$ is application specific and must be analyzed to identify the social network of the entities. So, a social network extraction model tries to construct a social network $G(V, A)$, where $V = \{v_1 = e_1, v_2 = e_2, v_3 = e_3, \ldots, v_n = e_n\}$ is the set of entities represented as the vertices of the graph G and A is the set of arcs or edges representing the relationships among the entities. Moreover, the social network extraction should be done in such a way that it best represents the information usage by the entities and is generalized with regard to the type of concepts the dataset D represents. Here, our focus is on the relationships that the edges represent so that the weight of an edge between two entities e_i and e_j is higher if their respective data usage is similar.

4 The Proposed Framework

To construct a social network for the entities $\{e_1, e_2, \ldots, e_n\}$, one can extract the important features from the dataset D_j associated with each entity. Then these features can be used to create a feature vector that will describe each entity. But care must be taken so that we only extract the useful information from the datasets and the dimensionality of the feature vectors representing the entities remains reasonable. Let us represent the feature vector related to entity e_j as $F^j = (w(f_1), w(f_2), \ldots, w(f_m))^T$, where $w(f_k)$ is the weight of the k-th feature, f_k in entity e_j and superscript T stands for vector transpose. So, the similarity between entity e_i and entity e_j can be calculated as the normalized dot product $\frac{F^i \cdot F^j}{\|F^i\| \times \|F^j\|}$ of their feature vectors. Other similarity measures like the Euclidean distance can also be used. Once we have the measure of distance and similarity, we can use one of the standard community extraction techniques to identify the communities.

Our work described in this chapter is based on the argument that frequent closed patterns can be used to generate the features that represent most of the useful information about an entity and hence can be used to construct the social networks. Using frequent closed patterns also helps to keep the dimensionality of the feature vectors to a reasonable size. Further reduction of the features is possible by filtering out features that do not satisfy certain threshold after computing the entropy of the features.

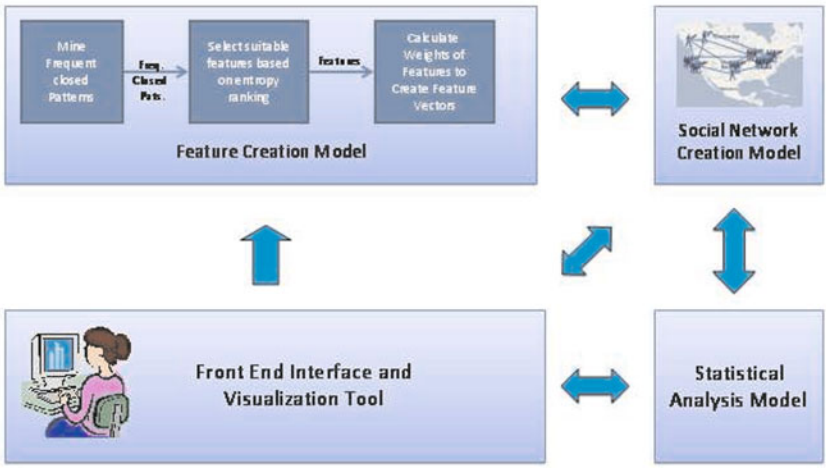


Fig. 1: The Proposed Framework.

Figure 1 presents the general block diagram of the social network analysis methodology proposed in this chapter. It has a feature extraction model that uses frequent closed patterns to create a feature set of reasonable size. Social networks can be constructed from the extracted features by invoking the social network creation model. Statistical analysis and social network visualization tools can be used to provide feedbacks to the feature selection and network creation steps. Each of the components shown in Figure 1 is described in brief in the rest of this section.

4.1 Feature Extraction Model

In order to best represent the user connectivity with the data, this model tries to create a feature vector to represent a user or entity. Here an entity may represent an individual, a set of individuals, a branch of a store, etc, depending on the goal of the organization. It is expected that each entity e_j is associated with a dataset D_j. Each dataset D_j is first converted into

a transactional dataset D_j^{tran}, where each transaction represents a user task or a set of user tasks, and each task represents a set of database accesses in D_j. Here, $D^{tran} = \cup_j D_j^{tran}, \forall j \in \{1..n\}$ is the transactional dataset for all the entities. This processing step of the database conversion is application specific; an example of this is given in Section 6.2 when we discuss the experimental results for the Enron email dataset.

Once the above database processing is done, the frequent closed patterns can be identified using one of the prominent frequent closed pattern mining techniques like CHARM [30]. Let the set of all frequent closed patterns be $P = \{P_i | P_i$ is a closed pattern and $sup(P_i) \geq minsup$ in $D^{tran}\}$, where *minsup* is a user specified threshold. Let us also define the closed pattern coverage, n_j of the Database D_j^{tran} as the number of all frequent closed patterns supported by the dataset D_j^{tran}. Formally,

$$n_j = |\{P_i \in P | P_i \text{ is frequent in } D_j^{tran}\}|,$$

where $||$ denotes the cardinality of a set. Following a methodology similar to the one described in [3], the entropy for the frequent closed pattern P_i can be defined as-

$$Entropy(P_i) = \sum_{D_j^{tran} \text{ s.t. } P_i \text{ is freq in } D_j^{tran}} \left[-\frac{1}{n_j} \ln \frac{1}{n_j} \right] \quad (1)$$

Algorithm 3 Steps of Feature Extraction

Preprocess: preprocess the given databases, $D_j, \forall j \in \{1..n\}$ & D to convert them into transactional databases, $D_j^{tran}, \forall j \in \{1..n\}$ & D^{tran}

Mine Frequent Closed Patterns: mine frequent closed patterns, P from D^{tran} (for a given support threshold) using one of the prominent frequent closed pattern mining techniques.

Calculate Closed Pattern Coverage: calculate the closed pattern coverage, n_j of the database $D_j^{tran}, \forall j \in \{1..n\}$.

Calculate Entropy: calculate the entropy of each of the closed patterns in P using Equation 1.

Rank: sort and rank the closed patterns based on their entropy.

Select: based on a user specified threshold, select the top few features as the final feature set.

Calculate Weight: for each entity, calculate the weight of each of the selected features using Equation 2.

The main motivation behind applying the entropy measure is that it allows us to identify frequent closed patterns for which the information content is higher. The set of frequent closed patterns are then sorted in ascending order and ranked according to their entropy values. Only the top few closed patterns are selected as features (based on a user specified threshold) to keep the number of features to a reasonable size. But ranking them using entropy ensures that only the best features are selected. As a result, we get a set of frequent closed patterns that make the features of the feature vectors (the

feature vector for entity e_j is, $F^j = (w(f_1), w(f_2), \ldots, w(f_m))^{tran}$, where $w(f_k)$ is the weight of the k-th feature, f_k in entity e_j). The weights of each feature is calculated using the following formula,

$$w^{D_j^{tran}}(f_k) = \frac{sup^{D_j^{tran}}(f_k)}{sup^{D^{tran}}(f_k)}, \tag{2}$$

where $w^{D_j^{tran}}(f_k)$ is the weight of the feature k for entity e_j, $sup^{D_j^{tran}}(f_k)$ is the frequency of feature f_k across the dataset D_j^{tran} of entity e_j, and $sup^{D^{tran}}(f_k)$ is the frequency of f_k across the dataset D^{tran} of all the entities E. Steps of the above feature extraction model are summarized in Algorithm 3.

4.2 Social Network Creation Model

The feature vectors created by the feature vector creation model serve in calculating the weights of each link or edge of the social network to be constructed. Here the weight of each link is inversely proportional to the Euclidian distance between the two feature vectors of the two entities that the edge connects. Once the social network is created, this model also helps in identifying the community structures of the particular social network created.

4.3 Statistical Analysis Model

This model will provide all the statistical analysis related to the social network created in the previous model. Tools like UCINET [27] or JUNG [10] can help the data miner with centrality measures, page rank measures, clustering, decomposition and other elementary graph theory measures; permutation-based statistical analysis methods help in analyzing the constructed social networks. Moreover, based on the feedback from the analysis, the data miner can adjust the feature vectors in order to achieve from the given dataset a refined social network structure; the latter structure is expected to represent the dataset in a better way.

4.4 Visualization Model

Tools like UCINET and JUNG can also be used to visualize the social network constructed by the social network creation model. Based on the visualization of the network, and the statistical measures collected at the previous phase,

the data miner provides feedback to the feature creation model in order to adjust the features appropriately.

Table 1: Behavioral Scenario

Individual Deviations	Community Deviations	Behavior Profile
low	low	User is consistent with both individual and community behavior profiles
low	high	User is consistent with individual profile but drifting away from community profile
high	low	User is consistent with community profile but drifting away from individual profile
high	high	User is drifting away from both individual and community profiles

5 Dynamic Behavior Analysis

In this section, based on the concept developed by Wan *et. al.* [28], we propose a methodology to dynamically analyze the behavior of individual actors over time. This is a necessary analysis step as it can help us to identify abnormal user behaviors as a result of hacking, intruder attack, or other unusual events like negotiations or deceptions [11] in an organization, an economic disruption in the market, an epidemic in an area, etc. This analyzing process is a part of the statistical analysis model where other statistical analysis is done and the social network is used as input to the model. By analyzing the social network produced by the social network creation model, we can identify how a user/entity is deviating from its usual behavior and the community behavior where it belongs [28]. According to [28] deviations only from individual normal behavior or only from community behavior do not necessarily depict an abnormal event. To detect an abnormal event or anomaly, we should consider the two kinds of deviations. The four possible scenario based on the individual and community behavioral deviations are shown in Table 1.

Traditionally, deviations are calculated as Euclidean distances. However, we argue that Euclidean distance alone does not provide us with the signifi-

cance of the distance value. Hence, a different kind of distance measure, the Mahalanobis distance [15] is used where the correlations of the sample points, hence how the points are spread over space are considered [5]. The Mahalanobis distance is considered as one of the most suitable distance measures for the multivariate data analysis [5].

Let us assume that each data accessed by an entity or a user is associated with a time period T_k and there are m such time periods, $T_1, T_2, \ldots, T_m$. For the purpose of extracting individual and community behavioral profiles, we divide the whole data into two sets: training period, $T_{train} = \{T_1, \ldots T_k\}$ and test period, $T_{test} = \{T_{k+1}, \ldots T_m\}$ for $k \in \{1..m\}$. For the training period, T_{train} each entity e_j's individual profile center μ_j and covariance matrix $\sum_j$ are computed. Then, individual deviation for time period $t \in T_{test}$ and entity e_j is calculated as the Mahalanobis distance.

$$Ind.Deviation, d_I = \sqrt{\left(F_t^j - \mu_j\right)^T \sum_j^{-1} \left(F_t^j - \mu_j\right)}, \tag{3}$$

where F_t^j is the feature vector for entity e_j at time period t. Next, based on the data set of the training period, community structures in the social network are identified. Let us assume there are n_c number of clusters, $C_1, C_2, \ldots C_{n_c}$ that categorize the entity dataset. Let us also assume that, μ_{C_i} and $\sum_{C_i}$ are the center and covariance matrix of cluster C_i for the training period, T_{train}. Then, the community/cluster deviation for a time period $t \in T_{test}$ and for an entity e_j belonging to cluster C_i is calculated as the Mahalanobis distance -

$$ClusterDeviation, d_{C_i} = \sqrt{\left(F_t^j - \mu_{C_i}\right)^T \sum_{C_i}^{-1} \left(F_t^j - \mu_{C_i}\right)} \tag{4}$$

Table 2: The entities and their corresponding Euclidean distances

	e1	**e2**	**e3**	**e4**	**e5**	**e6**	**e7**	**e8**	**e9**
e1	0.00	0.16	0.14	0.89	0.92	0.89	1.07	1.06	1.06
e2	0.10	0.00	0.12	0.84	0.86	0.84	1.03	1.02	1.01
e3	0.14	0.12	0.00	0.88	0.90	0.88	1.06	1.05	1.05
e4	0.89	0.84	0.88	0.00	0.07	0.00	0.74	0.73	0.72
e5	0.92	0.86	0.90	0.07	0.00	0.07	0.77	0.76	0.75
e6	0.89	0.84	0.88	0.00	0.07	0.00	0.74	0.73	0.72
e7	1.07	1.03	1.06	0.74	0.77	0.74	0.00	0.09	0.16
e8	1.06	1.02	1.05	0.73	0.76	0.73	0.09	0.00	0.08
e9	1.06	1.01	1.05	0.72	0.75	0.72	0.16	0.08	0.00

6 Experimental Results

We have tested our method on two data sets. The first one is a synthetic dataset that we have generated to test a hypothetical scenario where there are three groups of data users; each group of users consists of three entities that use datasets of similar characteristics. Our target is to see whether we can identify the original three groups of users based on the social network that we extract using the proposed methodology. The outcome from this initial test will be a first check for the validity of the developed methodology. In case the outcome is positive, then it will be a push forward to conduct the testing using larger and real datasets like the Enron-email dataset. Accordingly, we have also tested the proposed methodology to extract social network from the Enron e-mail dataset. Finally, the proposed feature vector construction method was used to analyze the dynamic behavior of each e-mail user. This dynamic behavior analysis can help to identify possible important events like change in user's behavior or e-mail account fraudulence, hacking, etc.

Table 3: The Enron E-mail users and their corresponding Euclidean distances.

	buy	dean	ermis	jones	kaminski	keavey	lokey	may	sager	saibi	salisbury	shakleton	thomas	whalley	ybarbo
buy	0.00	0.65	0.57	0.26	0.43	0.41	0.43	0.35	0.32	0.36	0.25	0.22	0.65	0.60	0.59
dean	0.65	0.00	0.13	0.50	0.28	0.50	0.27	0.68	0.40	0.44	0.73	0.64	0.08	0.10	0.13
ermis	0.57	0.13	0.00	0.44	0.22	0.44	0.21	0.61	0.33	0.38	0.65	0.56	0.15	0.14	0.16
jones	0.26	0.50	0.44	0.00	0.27	0.35	0.29	0.38	0.19	0.26	0.36	0.21	0.50	0.47	0.44
kamiski	0.43	0.28	0.22	0.27	0.00	0.31	0.16	0.47	0.17	0.28	0.51	0.39	0.28	0.25	0.25
keavey	0.41	0.50	0.44	0.35	0.31	0.00	0.38	0.25	0.30	0.41	0.45	0.38	0.51	0.47	0.50
lokey	0.43	0.27	0.21	0.29	0.16	0.38	0.00	0.50	0.22	0.25	0.52	0.41	0.27	0.25	0.24
may	0.35	0.68	0.61	0.38	0.47	0.25	0.50	0.00	0.40	0.45	0.35	0.33	0.69	0.65	0.67
sager	0.32	0.40	0.33	0.19	0.17	0.30	0.22	0.40	0.00	0.25	0.44	0.28	0.40	0.36	0.36
saibi	0.36	0.44	0.38	0.26	0.28	0.41	0.25	0.45	0.25	0.00	0.45	0.34	0.43	0.41	0.41
salisbury	0.25	0.73	0.65	0.36	0.51	0.45	0.52	0.35	0.44	0.45	0.00	0.30	0.75	0.70	0.70
shackleton	0.22	0.64	0.56	0.21	0.39	0.38	0.41	0.33	0.28	0.34	0.30	0.00	0.63	0.60	0.59
thomas	0.65	0.08	0.15	0.50	0.28	0.51	0.27	0.69	0.40	0.43	0.75	0.63	0.00	0.09	0.13
whalley	0.60	0.10	0.14	0.47	0.25	0.47	0.25	0.65	0.36	0.41	0.70	0.60	0.09	0.00	0.11
ybarbo	0.59	0.13	0.16	0.44	0.25	0.50	0.24	0.67	0.36	0.41	0.70	0.59	0.13	0.11	0.00

6.1 Synthetic Dataset

In this section, the proposed social network extraction model is tested for synthetically generated set of entities and their corresponding data repositories. First a synthetic data generator IBM Quest [2] is used to generate three sets of transactional databases of size 30K transactions each; the average

transaction size is 40 items and the total number of items in the database is 1000. The three sets of data were generated in three separate runs of IBM Quest so that each dataset contains transactions of similar properties. Next, each of the datasets is further partitioned into three more sets of size 10K transactions each, where each of the latter smaller datasets is assigned to an entity. Thus, 9 entities are created and each entity is associated with a 10k transactional dataset. Here, the datasets of entities e_{3n-2}, e_{3n-1}, e_{3n} (for $n = 1 \ldots 3$) come from a single dataset of 30K transactions, and hence possess similar properties. The rationale behind such a data generation method is that the set of entities e_{3n-2}, e_{3n-1}, e_{3n} are likely to fall into a single group forming a community or cluster.

All the datasets of individual entities are combined to create a 90K transactional dataset D. Frequent closed patterns are mined from this dataset using the closed pattern mining implementation as described in [9] with an absolute support threshold of 200. About 900 frequent closed patterns were identified using this method. Then, the entropy of all the frequent closed patterns was calculated. Based on the entropy ranking, 11 features were selected that possessed the most information contents. These features are {318, 170}, {585, 779}, {589, 886, 779, 813}, {577, 759, 188}, {339, 135, 507, 59}, {336, 63, 258, 130}, {238, 451, 575}, {40, 179, 693}, {253, 225, 533}, {524, 331, 701}, {623, 264, 640}.

Table 2 represents a matrix for each of the 9 entities, where each cell in the matrix represents the Euclidean distances between entity e_i and e_j with respect to their feature vectors, where i is the row number and j is the column number. We can observe from this matrix that the entities that belong to the same group/cluster have low Euclidean distances among themselves; whereas the entities that belong to different groups/clusters have high Euclidean distances among themselves. Low Euclidean distance values in the matrix are colored with red, green and blue colors to represent each of the three groups/communities, respectively. So, the useful grouping information is successfully portrayed by the proposed social network extraction method, where features are generated using frequent closed patterns.

6.2 Enron E-mail Dataset

For this experimental setup, we have used the Enron e-mail data [6] which has been publicly made available for researchers and is being commonly used as a benchmark dataset. This dataset contains 500,000 e-mail messages exchanged by 150 Enron employees. For our analysis, we only considered the e-mails that are found in the inbox and the people who have more than 1000 e-mails in the inbox.

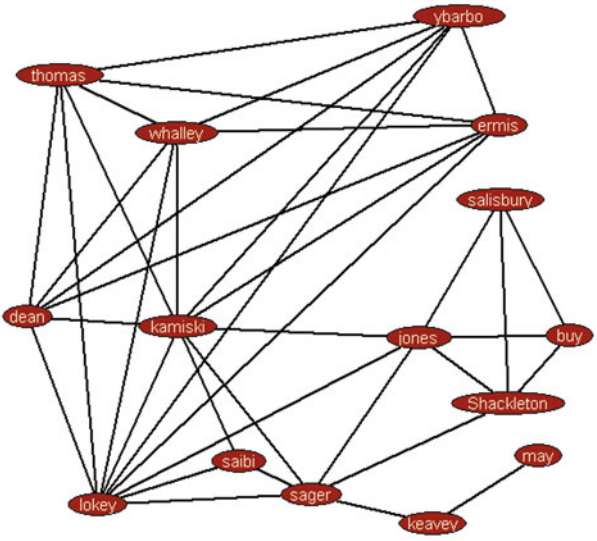


Fig. 2: The Enron E-mail users and their social network based on e-mail usage.

6.2.1 Data Processing and Network Extraction

From the 158 Enron e-mail user set, we have chosen a limited number of 15 users randomly in order to present them clearly in this manuscript. From each user's inbox, we have chosen 1000 e-mails randomly that makes the e-mail dataset for the corresponding user. Each e-mail is parsed to identify the stem words which make the itemset. Moreover, e-mail addresses inside the e-mails are identified as items as well. The stem words appearing in the subject line of the e-mails are also considered as items. These items appearing in a single e-mail are considered as a transaction. This way, for each user we construct a transactional database of items that appear in user e-mails. Here, each transactional database consists of 1000 e-mail transactions. From these transactional databases, we identify the global frequent closed itemsets (corresponding to a support of 10); then, based on entropy ranking, we chose the top 100 closed itemsets as our feature set. Table 3 presents the 15 Enron e-mail users and their corresponding Euclidean distances based on the E-mails in their inbox. Here, low distance means the users are related and high distance means the users are relatively unrelated.

Figure 2 shows the social network representation of the data reported in Table 3. In this case, we only considered the edges if the Euclidean distance between two entities is less than a user specified threshold of 0.30. Here, it

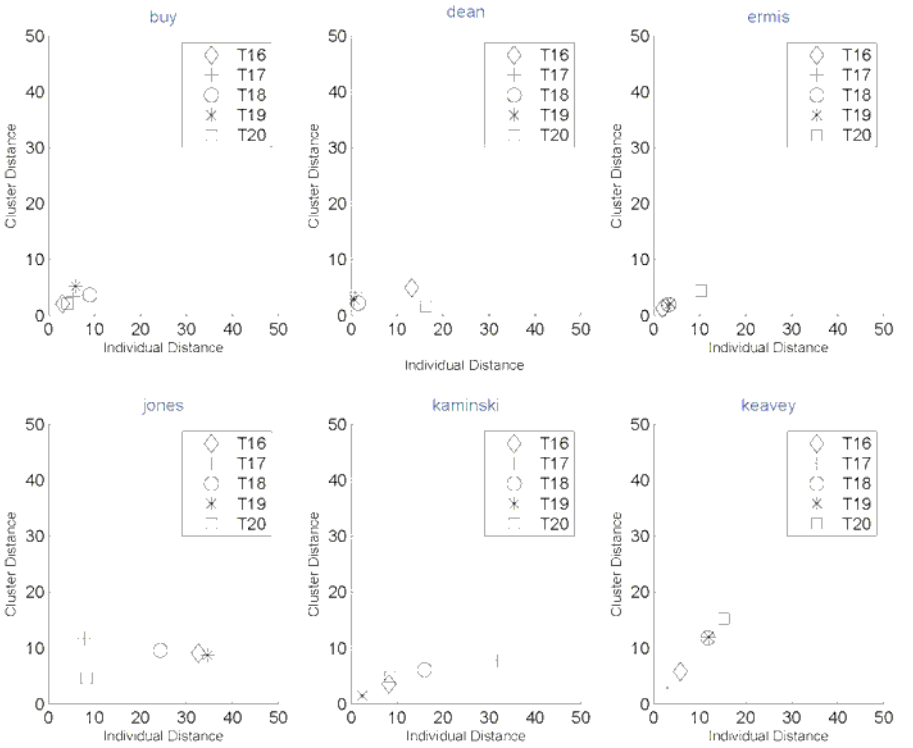


Fig. 3 (a) – Scatter Graph of Mahalanobis Distances.

is clearly depicted that some users are related to many other users; whereas some users are isolated. For example, *kaminski* is related to 8 other users and on the other hand, *may* is only related to *keavey*.

6.2.2 Dynamic User Behavior

For this analysis, we only consider the emails in the duration from July 1999 to December 2002. First, we arbitrarily and equally divided the period into 20 intervals of equal lengths. For each of the 20 periods, we collected the emails in each period for each user. Next, using the proposed methodology, we constructed the normalized feature vectors for each user and for each of the 20 periods. The first 15 periods ($T_1 \sim T_{15}$) comprise the training dataset; while the final 5 periods ($T_{16} \sim T_{20}$) comprise the test dataset.

In the next phase, a clustering analysis technique is used to identify the most prominent clusters (of email users) in the training period. We have

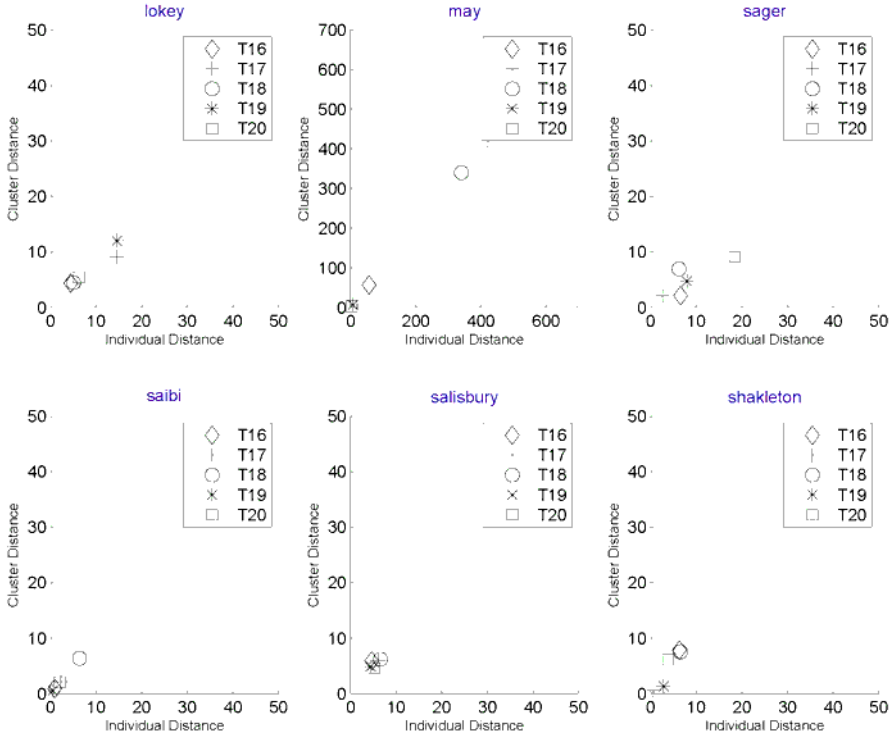


Fig. 3 (b) – Scatter Graph of Mahalanobis Distances.

Table 4: The Clusters of Enron E-mail.

Cluster Id	Enron Users
1	saibi
2	buy, salisbury, shakleton
3	dean, ermis, jones, kaminski, lokey, sager, thomas, whalley, ybarbo
4	keavey
5	may

used the hierarchical clustering technique provided in *'matlab'* and found that there are five prominent clusters. The discovered five clusters are shown in Table 4.

This result, not surprisingly, is consistent with our findings described in Section 6.2.1. From Figure 2, it is obvious that there are three isolated users which are connected to either only a single user or only a couple of users.

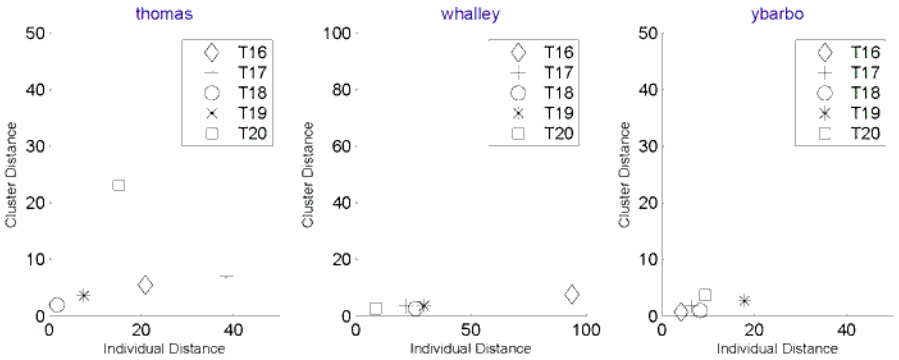


Fig. 3: **(c)** – Scatter Graph of Mahalanobis Distances.

These three users namely *may*, *keavey*, and *saibi* are outliers and thus formed single entry clusters. The three users *buy*, *shakleton*, and *salisbury* are connected to each other and to the user *jones*. So, they formed a cluster. The rest of the users are similar and are connected to most of the other users. Hence, they formed another cluster.

Once we identify the clusters from the training dataset, we compute each individual user's individual Mahalanobis distance (using Equation 3) and the cluster Mahalanobis distance (using Equation 4) for each of the test periods $T_{16} \sim T_{20}$. Motivated by and adapted the method developed by Wan *et. al.* [28], we next plotted the scattered diagram of the Mahalanobis distances for each user and for each test period. The result is shown in Figure 3. Here, high individual Mahalanobis distance means the user is drifting away from the individual email profile and high cluster Mahalanobis distance means the user is drifting away from the cluster profile. As mentioned in Table 1, there can be four possible scenarios based on the individual and cluster Mahalanobis distances. The fourth row in Table 4 has the most interesting scenario and may imply an anomaly.

In Figure 3, the user *whalley* deviates from his individual email profile in time period T_{16}. This may mean that a new profile for the user is seen; and this information is consistent with *whalley*'s group profile. Similar behavior is exhibited by *jones* in the time periods $T_{16}, T_{18}, and T_{19}$; similar behavior is also exhibited by *kaminski* in the time period T_{17}; and it is as well exhibited by *thomas* in the time period T_{17}. For the three users *may*, *saibi*, and *keavey*, the points in the scatter plots are linear because in these cases the users are outliers (so part of a cluster of one member). The user *thomas* deviates from both individual and cluster profiles in the time period T_{20}; this deviation may imply the need for further investigation of this case because it is not

the user's normal behavior. However, on the other hand, this deviation is not by a significant amount; so it can be a false alarm. In the time period T_{18}, *may* exhibits significant deviation from the individual profile and this requires further investigation. As a result, this brief analysis should be enough to highlight the importance of studying the dynamic behavior of the actors in the social network. However, such a study requires the availability of data for a number of periods in order to allow for the analysis to expand and cover the periods consecutively.

7 Conclusion and Future Work

This chapter presented a social network modelling technique that takes as input the data to be analyzed for constructing a social network and maps it into a new space by mining frequent closed patterns. Using the produced frequent closed patterns, we are able to create a useful set of features to represent an entity that describes the connection of the entity to data in a reasonable way. Moreover, entropies of the collected features are used to find a reasonable set of features while keeping the information content high. The feature vectors created in this way facilitate a social network extraction model that maintains all the community structure information. The results presented in this chapter also verify the above conjecture both for the synthetic and real datasets. Currently, we are investigating how the proposed approach could scale to large scale network consisting of thousands (even millions) of actors. For large scale networks, we may consider groups of similar users as a single entity and test how these groups of users may relate to each other. Effective grouping of users should be the key here. Finally, we will apply the same methodology described in this chapter on financial data. The financial data becomes available incrementally over time and hence it is very suitable and convenient to be considered for analysis to discover its dynamic behavior by employing the methodology described in this chapter.

References

1. R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases," in *SIGMOD Conference*, 1993, pp. 207–216.
2. R. Agrawal, M. Mehta, J. C. Shafer, R. Srikant, A. Arning, and T. Bollinger, "The quest data mining system," in *KDD*, 1996, pp. 244–249.
3. F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering," in *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.* New York, NY, USA: ACM, 2002, pp. 436–442.
4. G. Carenini, R. T. Ng, and X. Zhou, "Summarizing email conversations with clue words," in *WWW '07: Proceedings of the 16th international conference on World Wide Web.* New York, NY, USA: ACM, 2007, pp. 91–100.

5. M. R. De, J.-R. D., and M. D. L., "The mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 1, January.
6. "Enron email dataset," http://www-2.cs.cmu.edu/$\sim$enron/, retrieved Sep 03, 2009.
7. G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, "Self-organization and identification of web communities," *Computer*, vol. 35, no. 3, pp. 66–71, 2002.
8. M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *PNAS*, vol. 99, no. 12, pp. 7821–7826, June 2002.
9. G. Grahne and J. Zhu, "Efficiently using prefix-trees in mining frequent itemsets," in *FIMI*, 2003.
10. "Jung: Java universal network/graph framework," http://jung.sourceforge.net, retrieved Sep 03, 2009.
11. P. S. Keila and D. B. Skillicorn, "Detecting unusual email communication," in *CASCON '05: Proceedings of the 2005 conference of the Centre for Advanced Studies on Collaborative research*. IBM Press, 2005, pp. 117–125.
12. B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *Bell System Tech. Journal*, vol. 49, pp. 291–307, February 1970.
13. F. M. Khan, T. A. Fisher, L. Shuler, T. Wu, and W. M. Pottenger, "Mining chat-room conversations for social and semantic interactions," Lehigh University, Bethlehem, PA, Tech. Rep. LU-CSE-02-011, 2002.
14. J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. of Fifth Berkeley Symp. on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.
15. P. C. Mahalanobis, "On the generalised distance in statistics," in *Proceedings National Institute of Science, India*, vol. 2, no. 1, April 1936, pp. 49–55. [Online]. Available: http://ir.isical.ac.in/dspace/handle/1/1268
16. N. Matsumura, D. E. Goldberg, and X. Llorà, "Mining directed social network from message board," in *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*. New York, NY, USA: ACM, 2005, pp. 1092–1093.
17. P. Mika, "Bootstrapping the foaf-web: An experiment in social network mining," in *Proc. of the 1st Workshop Friend of a Friend, Social Networking and the Semantic Web*, Galway, Ireland, 2004, pp. 1–2.
18. E. Minkov, W. W. Cohen, and A. Y. Ng, "Contextual search and name disambiguation in email using graphs," in *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2006, pp. 27–34.
19. G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, June 2005.
20. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering frequent closed itemsets for association rules," in *ICDT '99: Proceedings of the 7th International Conference on Database Theory*. London, UK: Springer-Verlag, 1999, pp. 398–416.
21. F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *PNAS*, vol. 101, no. 9, pp. 2658–2663, March 2004.
22. J. Scott, *Social Network Analysis: A Handbook*, 2nd ed. Sage Publications, 2000.
23. J. Shetty and J. Adibi, "Discovering important nodes through graph entropy the case of enron email database," in *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*. New York, NY, USA: ACM, 2005, pp. 74–81.
24. S. Staab, P. Domingos, P. Mika, J. Golbeck, L. Ding, T. Finin, A. Joshi, A. Nowak, and R. R. Vallacher, "Social networks applied," *IEEE Intelligent Systems*, vol. 20, no. 1, pp. 80–93, 2005.
25. S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, March 2001.

26. J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, "Email as spectroscopy: automated discovery of community structure within organizations," pp. 81–96, 2003.
27. "Ucinet: Social network analysis software," http://www.analytictech.com/ucinet/, retrieved Sep 03, 2009.
28. X. Wan, E. Milios, N. Kalyaniwalla, and J. Janssen, "Link-based event detection in email communication networks," in *SAC '09: Proceedings of the 2009 ACM symposium on Applied Computing*. New York, NY, USA: ACM, 2009, pp. 1506–1510.
29. H. Yu, D. Searsmith, X. Li, and J. Han, "Scalable construction of topic directory with nonparametric closed termset mining," in *ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2004, pp. 563–566.
30. M. J. Zaki and C.-J. Hsiao, "Charm: An efficient algorithm for closed itemset mining," in *SDM*, 2002.

Part II
Dynamics in Social Network Models

Visualisation of the Dynamics for Longitudinal Analysis of Computer-mediated Social Networks-concept and Exemplary Cases

Andreas Harrer, Sam Zeini, and Sabrina Ziebarth

Abstract In this paper we will demonstrate the potential of processing and visualising the dynamics of computer-mediated communities by means of Social Network Analysis. According to the fact that computer-mediated community systems are manifested also as structured data, we use data structures like e-mail, discussion boards, and bibliography sources for an automatic transformation into social network data formats. The paper will demonstrate a 3-dimensional visualisation of two cases: the first presents an author community based on bibliography data converted into GraphML. Based on this dataset we visualise publications networks with a tool called Weaver, which is developed in our research group. According to Lothar Krempel's algorithm, Weaver uses the first two dimensions to embed the network structure within a common solution space. The third dimension is used for representing the time axis and thus the dynamics of co-authorship relations. The second case describes recent research in open source communities and highlights how our visualization approach can be used as a complement to more traditional approaches, such as content analysis and statistics based on specific SNA indices.

Andreas Harrer,
Katholische Universität Eichstätt-Ingolstadt,
e-mail: andreas.harrer@ku-eichstaett.de

Sam Zeini,
Universität Duisburg-Essen, Germany,
e-mail: zeini@collide.info

Sabrina Ziebarth,
Universität Duisburg-Essen, Germany, Germany,
e-mail: ziebarth@collide.info

1 Introduction

Social Network Analysis (SNA; for an overview please see: [15]) is becoming more and more interesting to the domains of computer science related to communities. Besides the body of work of computer scientists in the graph drawing community[1] (where we were inspired by the work of Ulrik Brandes and Dorothea Wagner), we believe that SNA will once be established on par with statistical analysis as well as with qualitative approaches in the interdisciplinary fields of "Computer Supported Cooperative Work" (CSCW) and "Computer Supported Collaborative Learning" (CSCL), where empirical grounding is broadly common to the community. As an example of current development for CSCW one can mention the workshop on SNA at the international CSCW conference in the year 2004 [2]. In the field of CSCL research using SNA was presented to a broad audience at the international conference on the learning sciences by Palonen and Hakkarainen [11] and the international CSCL conference by Reffay and Chanier [12] and is now also published in international journals within the community (e. g. [9] or [5]). We identify two major reasons for attractiveness of the SNA approach to these two research domains, which are related to each other. The first reason is that wherever groups or communities are mediated through computer systems, a representation of the community is also coexistent in the computer systems as structured data. This "natural data" often can be transformed easily into social network data, once we assume the networks being defined as closed networks by the fact of user management implicated by the systems. Such systems can be well known basic technical support approaches like mailing lists or discussion forums. They also can be advanced cooperation tools like the BSCW system[3] or advanced collaborative learning environments like shared workspace systems in combination with learning object repositories (e. g. [6]). The basic concepts on which we build the data transformations will be a part of this paper.

The second reason which is related to the operational availability of data is related to the possibilities of visualisation provided by concepts and techniques of Social Network Analysis. According to Krempel [7] advanced visualisation techniques should be able to impart complex knowledge in a very efficient way. In the area of CSCL the potential of awareness through visualisation of collaborative and social aspects is discussed for several target groups:

Researchers can use these visualisations as well as the related SNA indices as means to support other methods of analysis [5][9] since triangulation research design is common ground within the interdisciplinary CSCL field. Teachers are enabled to understand the group structures in their computer

[1] www.graphdrawing.org

[2] http://projects.ischool.washington.edu/mcdonald/cscw04/

[3] http://www.bscw.de/english/index.html

supported classes and courses (e. g. for school classes using discussion forums or blogs and university courses with blended scenarios) and potentially can use this information to guide and advice the students, e. g. when participation of specific students is extraordinarily high or low. Finally, *students* could use the visual feedback for self reflection and self regulation in reaction to the information transported (cp. [13]). Other potential scenarios related to work in the CSCL are situations in further education, where social networks are critical to the aspects of development of competencies on the job. Here such awareness support systems can provide orientation within networks, which are often complex.

A third aspect which emerges from both reasons described above is that the data captured from the collaborative systems contains additional information. In our cases the timestamps captured by these systems are used to include the dynamics of the networks in the representation. For these reasons we developed a toolset of transformation concepts and tools that we describe in the next section as well as techniques for visualisations to support the mentioned user groups and that are tailorable for their specific needs presented in section 3.

2 Data Collection and Processing

The data used in our network analyses is - in the classification of Wassermann & Faust [15] - mainly of the category "archival records", i.e. compiled in formal and structured documents that can be parsed with the help of computer programs. Our initial research interest [4] was the support of asynchronous collaboration in thread-based discussion forums. The data that is created there provides well-defined connections between the contributions written by the network actors. Formally, these networks of postings are connected with each other in a "refers-to" relation and each posting is created by a network actor. These networks can be represented graph-theoretically as two-mode networks. Depending on the complexity of the discussion forum, the posting entities in the network might have several additional attributes interesting for analyses, such as creation time, categorical information (e.g. speech act categories [1] that have been used by us in the SPREKON system [4]), intended readers, detailed text etc.

Building up on first analyses of these communication archives, we transferred our data formats and tools to other archival collaboration data, such as third-party web portals, mailing lists, log files from source code repositories, and biblio-graphical data.

All these data sources are mapped to graph-theoretical data structures, that provide a uniform XML-binding, and that can be used for a variety of data conversions and analysis techniques. Among these are typical operations, such as reductions from two-mode to one-mode networks [15], conversions to

third party formats such as UCINET, Pajek, or GraphML, and our own analysis and visualisation techniques, as the one we describe in this paper. An overview of the formal aspects of conversions between different network types can be seen in Fig. 1.

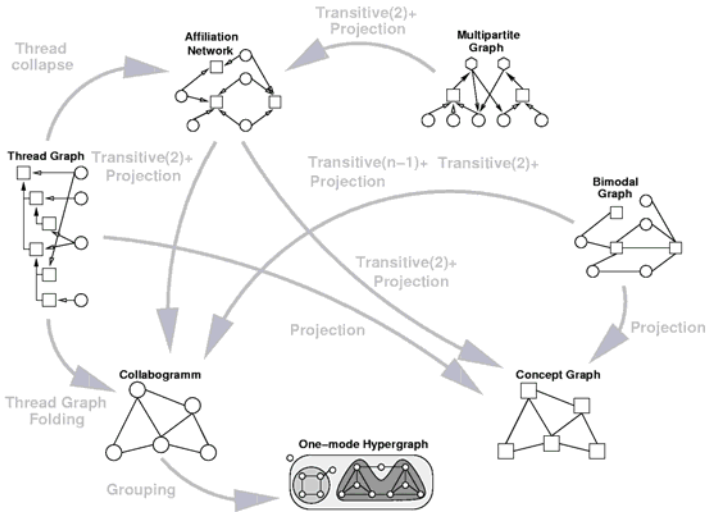


Fig. 1: Schema for formal network transformations

For the example cases in this paper we will focus on bibliographic networks and mailing lists: bibliographic networks consist of authors, publications, and potentially the conferences / journals / book series it is published in. This is a typical case of a multi-partite graph, that can be processed further to a bipartite author-publication network (comparable to an affiliation network) and finally to an author network (comparable to a sociogram or - as we call it in CSCL contexts - a collabogram). Other typical graph-theoretic transformations we support are based on thread graphs that occur naturally in discussion forums and mailing lists (see above) and that can be transformed to either strict person-topic networks (bi-partite type) or one-mode networks (either focussing on persons or on topics, similarly to the duality of affiliation networks). Grouping and partitioning of actors into subsets with similar properties or shared relations (this can be represented by a hyperedge or enriching the actor attributes) will also be discussed in this paper and the proposed visualisation approach presented in the next section is designed to specifically to support the user in interpreting these network structures.

3 Visualisation Approaches – Towards a Representation of Network Dynamics

In addition to powerful methods and algorithms for analysing social networks, it is often very useful to have a visualisation of the results which is as simple as possible. A very common visualisation of a social network is a sociogram, which contains actors as nodes layouted on a circle and their relationships as edges connecting the nodes. While this kind of visualisation seems to be self-explanatory it might also lead to misinterpretation. Although the nodes are randomly assigned to the circle and the lengths of the connecting lines have no meaning, users tend to interpret visual adjacencies as social adjacencies. Thus, the layout of the nodes is a central problem of network visualisation. Krempel [7] suggests an algorithm for arranging the nodes in simple solution spaces. A solution space defines a set of feasible positions for the nodes, e.g. coordinates on a circle. To optimise the layout of the network graph, edges should have as few crossings as possible and thus highly connected nodes should be placed close to the nodes they are connected with.

Unfortunately, the calculation of the optimal positions of the nodes would require testing all of their possible positions, which has high computational efforts. To avoid testing all permutations, Krempel (ibid.) arranges the nodes with the highest structural importance first and accepts that nodes which are structurally less important are arranged less optimal. Structural importance is defined by the centrality measures of the nodes, thus central nodes, which have either many direct or many short indirect links to other nodes, are placed with priority. An advantage of the use of simple solutions spaces is that nodes can be placed according to their properties in predefined regions of the visualisation. For example, each set of nodes belonging to one time interval could be arranged on one circle. Our Weaver application uses Krempel's algorithm (ibid.) for arranging the nodes according to their properties and the perspective the user is interested in, e.g. there are special views for group visualisation or multi-mode networks. Furthermore, Weaver allows the definition of the appearance of the nodes according to their properties in the form of rules, which can be saved and exchanged as rule sets related to specific type of networks. For example, in a two-mode network actors could be represented as filled circles and events as squares, and the nodes could have a size proportional to their importance and colour by defined other properties, such as formal rules or hierarchy in a group. From the user view this means, that he or she is able to perceive the properties of a node immediately and intuitively.

The consideration of the dynamics of networks over time has been discussed early in the SNA literature, as the well known example of the Sampson monastery data (cp. [10]) shows. Yet, the capturing of discrete snapshots of a social network for each interesting moment and the isolated diagrammatic representation according to the approaches mentioned in the previous chap-

ter does not properly support the reception and interpretation of network dynamics easily.

Some related work relies on animated transitions through the snapshots to integrate the temporal dimension (e. g. the commetrix system, cp. [14]). In our approach, we make use of the third dimension for the representation of dynamics: all interactions in our archival data contain timestamp information, that represents creation date, publication data etc. Thus it is possible to augment the typically two-dimensional sociogram representations with temporal information along a third dimension. Thus it is possible to augment the typically two-dimensional sociogram representations with temporal information along a third dimension, similarly to the approach in [3].

One possibility we explored for visualisation using a temporal dimension is the radial rendering with a "starting point" (i.e. the timestamp of the first interaction) and subsequent timestamps on concentric circles surrounding the previous ones. While this creates an intuitive grasp of the degree of activity at each time slice, the optimized graphical rendering of actors and their close collaborators is a difficult challenge with this approach. Even when each actor is represented by its own radius line in the diagram, the visual representation inherently would imply that relations between the same actors at later time stamps are more distant / less intense than at earlier time stamps. This might obscure or inhibit the intuitive interpretation of the diagram type.

In our proposed visualisation method, the integration of temporal dynamics is conducted as follows: Each network actor is represented with a timeline showing the timestamps her interactions took place at. These timelines are oriented for all actors along the third (Z) dimension. To optimize a graphical rendering of a "full community sociogram", all the actors within one timeslice are taken into account to define the layout in two dimensions (X-Y) according to the metrics of embedding into solution spaces[7]. Thus, when the whole community is of interest for analysis a flat X-Y view is chosen. For a conventional snapshot representation of the network community at a given moment, the filtering mechanism of our visualisation tool can be used using the timestamp of interest as the filter. For the integrated perception of the dynamics of the network, the "flat view" is rotated along the Z-axis, revealing both the individual actors' timelines and the different slices representing network interaction at one given moment. A detailed inspection of interesting aspects is further supported by the zooming mechanism and a filtering on specific actors. The next section presents some details about our analysis process with respect to data processing steps, used formats and manual interventions needed for data cleaning which is followed up by two case studies using this method.

4 Implementation

The data processing for our visualisation approach follows along a tool chain we developed in recent years: Archival data, for our scenario in this paper formal bibliographic data in the Bibtex-format, is parsed using a configurable parser generator provided by the Data-Multiplexer-Demultiplexer (DMD) application. This flexible approach allows us to handle different input sources such as mailinglists and newsgroups, discussion forums [4] and photo community web galleries, log files from source code repositories, as well as our own formats SPREKON [4] and CoNaVi (used by our Community Navigation Visualiser [8]). The parsed entries are mapped to our internal SPREKON data structure representing the publications, the authors and potentially the editors. This network can be exported as a multimode network or can be reduced to an author network in different formats, such as Pajek, UCINET, or for this purpose a specific GraphML dialect (http://graphml.graphdrawing.org/). Here it is also possible to export networks which contain the thread graphs representing the relations between the topics. The DMD application also allows the user to choose directed or undirected graph exports. For us, one of the important aspects of the internal handling of the thread graphs is the possibility to derive directed one-mode person networks from the multimodal communication networks. DMD uses an internal object structure which consists of a set of postings based on the SPREKON format and provides us the flexibility to extend DMD for new input and output formats.

One of the main problems we faced during converting large sets of electronic communication data is the possible redundancy of users e.g. a user is known by different email accounts. To solve this problem we introduced a user merge function in DMD which provides several filtering concepts (similarity, Levenshtein, etc.) to support the researcher while cleaning the noise of the source data. Last but not least the DMD application handles also the temporal information in the data transformation process. Here the user is able to choose between the exact timestamps or a slicing into discrete time points representing an interval of days, months, or years. Currently we support timeline information in the output of Pajek and GraphML files.

That GraphML file can be imported into our Weaver application and is best viewed then with our "temporal dynamics" view and an appropriate rendering style that gives specific visual attributes to actors based on the actor properties (e.g. actor size proportional to actor's centrality). According to [7] techniques exist to integrate structural properties of networks in the display. For this purpose we created the Weaver application, a 3D visualiser for social networks, which arranges and draws the nodes according to properties such as degree, centrality, or externally defined properties within a simple solution space. From the user view this means, that he or she is able to perceive the properties of a node immediately and intuitively (e.g. what is the most central topic in the network).

5 Example Cases for the Visualisation Method

To demonstrate the features of our visualisation method and our tool chain, we used the approach described above to analyse two case studies. In the first case we analysed a specific set of bibliographic information that was taken from a citation network of CSCL literature with special focus on computer-supported collaboration, interaction analysis, and communication networks. In the second case we analysed the communication and collaboration in the open source project OpenSimulator with a focus on innovative contributions made by users in the community.[4]

5.1 CSCL Citation Network

The raw data was represented in the Bibtex-format, as structured format containing information about authors, publication titles, type of publication, editorial information, publication date etc. Frequently bibliographic data contains multiple writings for the same persons: Sometimes the publisher's formats demand full names, sometimes only the "initial" of the first name, and some authors even vary their names across publications using or not using "middle initials" or double family names.

To identify persons with different writing across the publications we used the "user merge" feature of DMD and thus cleaned the data for further analyses. While we have features to visualise and analyse multi-mode networks, we will concentrate in this example on one-mode co-authorship networks created from this data. This source data was converted by the DMD tool to the GraphML format and - for comparison reasons - to the Pajek .net format. Fig 2 shows an isolated diagram produced by the Pajek application [2] with our time slice export. By navigating through the single slices the user has to interpret the dynamics of the network on his own.

The visualisation method we propose in this paper uses the GraphML output produced by the DMD. Using the temporal information in the third dimension, we get a representation that can be manipulated interactively by the user through changing the perspective in the 3-D coordinate system. A flat view of the 3-D visualisation (not shown for reason of space limits) would show the temporal axis from right to left. Each vertical set of authors represents the authoring network in a specific time interval. Actors that published at various intervals are represented using colour-highlighted "lifelines" symbolising their publication history. Actors with numerous co-authors are rendered in different colours than authors that publish with few co-authors, using the SNA properties and algorithms of k-core [2]. These augmentations

[4] www.opensimulator.org

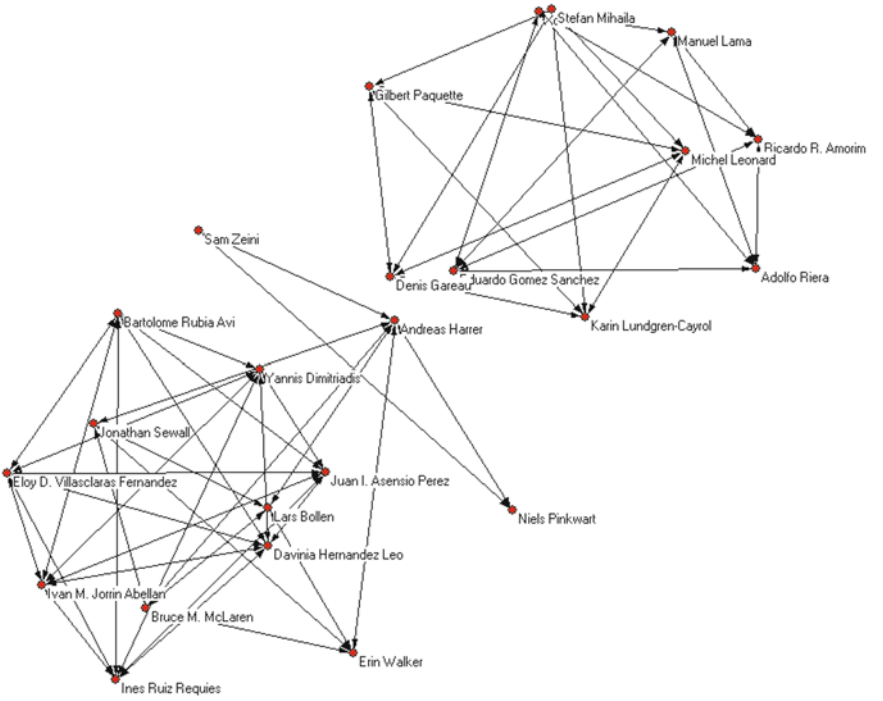


Fig. 2: Isolated time slice diagram of the author network

of the network by visual differentiation are supported flexibly in Weaver using node styles and edge styles that define rule sets for visual rendering.

The use of all three dimensions simultaneously can be seen in Fig. 3where the 3-D model is shifted slightly towards the third axis. While each time slice is still discernable, the information about the full author network at each timepoint is additionally visible. Because the authors' history (lifeline) is also visible, this perspective is well suited to get a first-glance impression of the network including the temporal dynamics. When the user is interested especially in zooming on specific periods of time, another feature of the Weaver application, a flexible filter mechanism can be used. A filter on the year 2005 produces the timeslice represented in Fig. 4, which is well comparable with the network view in Fig. 2.

Using combinations of filters, finer selections on arbitrary subsets of the whole dataset can be visualised flexibly in the same manner as in the previous figures (e.g. concentrating on the latest 5 years of publication in the shifted 3-D view). Filters can also be used on specific persons to visualise their publication histories in an isolated view or only with their close co-authors.

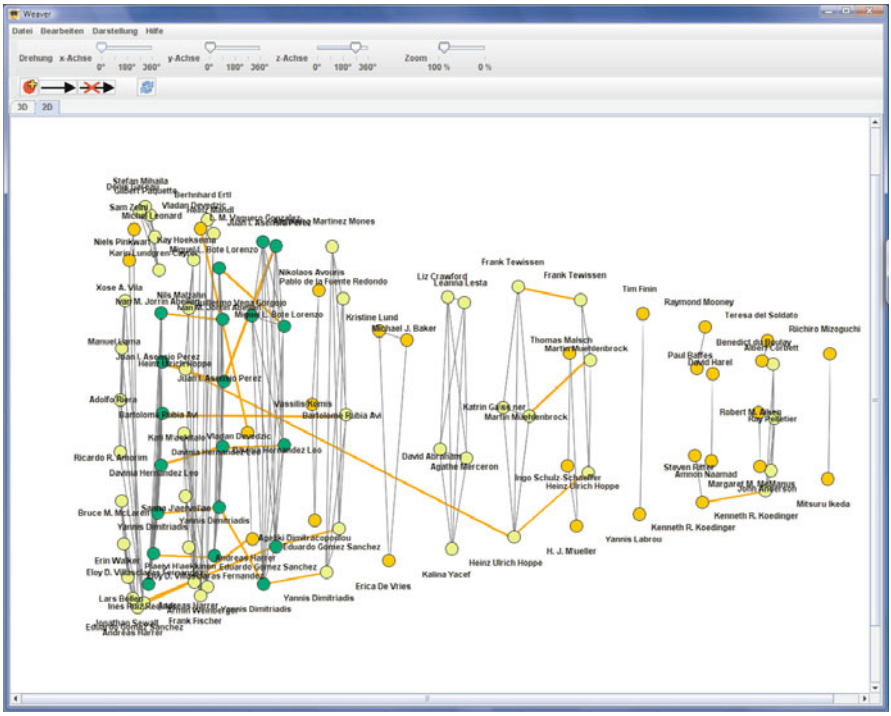


Fig. 3: Using all three network dimensions simultaneously

5.2 OpenSimulator Network

Within the scope of a current research project, we used the approached discussed in this paper to elaborate an alternative view to one case study which was focused on investigation. In general the research project is interested in open source innovation processes within business related communities and how these insights can be applied to the classical IT sector. OpenSimulator is a server platform for hosting virtual worlds (similar to Second Life). There is an active developer community and the user group is also lively.

One interesting point about open source projects is that not only the source code is publicly available. Most open source projects also cultivate an open culture of documentation. In this sense one can also get access to the archives of the mailing lists. The DMD tool enabled us to transform these mailing list archives into social network data. Also the source code archives contain the information about who worked on the code at which time, so we can also use them as an additional source for social network data.

For the purpose here we focused on a longitudinal analysis of the developer mailing list, the community mailing list and the SVN version management system log files from September 2007 to February 2009. The Developer mail-

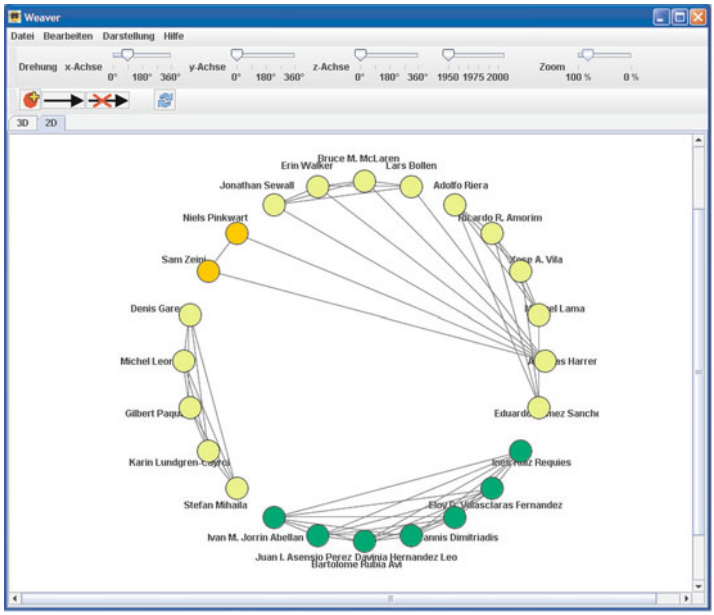


Fig. 4: Projection into one timeslice using a filter

ing list contains 198 users, 5505 mails referring to 1185 topics. The SVN data contains 32867 changes applied to 6012 objects, where the changes have been processed by 26 developers. The community mailing list contains 175 users discussing 634 topics in 1582 mails.

By just looking into the data we can directly get some first insights which are often typical for the organizational structure in open source projects. First of all we can observe, that there are just 26 persons holding committer permissions while there are 198 users communicating in the developer mailing list. This difference is interesting since it can explain one major aspect of the hierarchical situation given in the OpenSimulator project. By definition the source code of an open source project can be accessed and modified by everyone who is able to understand and work with the code. Usually open source licenses regulate the distribution of the modified code but this does not inevitably mean that the modified code will be applied to the original project by everyone. So in most projects there are fewer developers with committer rights then free developers. The loosely assigned developers usually send their modifications via the mailing list as patches which will be reviewed and applied by the developers with committer access to the source code management system. Users with committer rights, who are formally assigned with the role of a maintainer, usually are responsible for specific parts of an open source project. When someone often sends useful patches to the mailing list he or she receives an invitation to a maintainer position in the project. On

the other hand, good patches regulary pass through the mailing list without long discussions. They fit into the project and get applied fast.

This explains - by comparing the developer mailing list with the SVN network - that developers with high quality contributions / patches like Melanie T. are not very central in the mailing list (degree centrality) until they become maintainers and join the SVN network. This should also be taken into account as a possible bias while analyzing open source networks.

Comparing the structures of the three different networks (SVN, developer mailing list, community mailing list) the results show a 0,05% ($p = 0{,}012$) level significant middle strong correlation (Spearman) of users (0,592) between SVN activities and developer mailing list, while the top committer is not congruent with the top poster. There is also a significant ($p = 0{,}01$) middle strong correlation (0,640) between developer list and community list.

These first insights lead us to the assumption that events such as the patch example are important for the longitudinal analysis so we want to take a look at the degrees of the topics in the two-mode version of the dataset derived from the developer mailing list. For this purpose we cut off all topics with a degree centrality less than 100 from the network. While most of the high degree topics are often very technical discussions about details which can be identified quickly since the subjects of the mails are tagged, we can identify one topic with a degree of 107 where a user is suggesting essential changes. In this case the user Terry F. suggests to handle voicechat, textchat and friends list on the OpenSimulator client's side to allow users to carry their inventory from world to world and be able to monitor friends. He seems to be new to the list and has no advanced skills in programming so he comments his own request by the sentence "Please forgive me if this sounds a bit noobish". He receives positive feedback from core developers like from Charles K. with the words: "As I think about it, there will come a time when certain things are most appropriate to hold on the client side. Things such as clothes, tools, wallets, cellphones...". The suggestion of Terry induces an intense discussion where the team works on some new ideas and concepts.

Since this user makes a potential innovative contribution to the network we decide to target his further activities within the community. The k-core metrics [15, 2] computed for each project month shows that Terry F. didn't exist in the core regions of the network before his request in March 2008 (month 15 counted from January 2007). After his suggestion he achieves a position between core and periphery (core values 3 out of 5 for March in that month and holds a comparable position until August. This can also be seen at a glimpse in the visualisation of the central actors in Fig. 5. The actor in our focus (highlighted as a triangle that is sized approximately in double size compared to the other actors to make him clearly visible) appears with substantial core in month 15 and keeps this until month 19.

Focussing on him visually by using the filtering options of weaver we see some other potential important events on the timeline in Fig. 6. We can obviously detect the second situation where he contributes another innovative

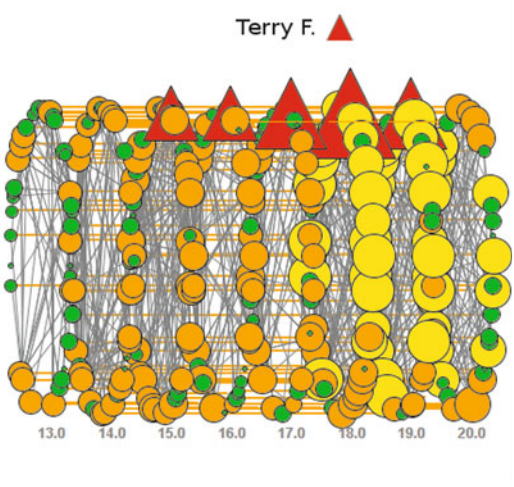


Fig. 5: Timeline visualisation (months 13 to 20) of the actors with highest core value

idea to the project by observing the number of interactions with other members of the community: In July 2008 (month 19) he suggests to configure the load balancing mechanism of OpenSim to act as a mirror for backups. Also in this case he receives an unreserved reaction from the community exemplified by a typical the comment from the developer Johan B. "I only want to know that this feature is useful for someone before I start coding".

By zooming in even more into the specific month July 2008 (M 19) and exploring the degree of interconnection using our k-cylic group arrangement we can clearly see in Fig. 7 that our focused actor is well connected at this time even though his absolute values on indegree and degree are only moderate (4 each, while other actors in the diagram have up to 20 in these values). All actors located in the circle in the upper left are connected with 3-cycles to other members of this subset and are scaled in size according to their indegree.

With the combination of classical techniques of Social Network Analysis with our visualization approach we were also able to dive immediately into a complex technical community and observe that new users like Terry can contribute to open source projects in an innovative way without having skills at an expert level. This kind of observation method with an orientation towards ethnographic methods, that we call E-Ethnography, uses Social Network techniques as an heuristic access to understand communities: The longitudinal visualization by means of the Weaver approach enables us to integrate the mathematical indices, such as k-core, smoothly into a story telling perspective as we hope we were able to demonstrate in the second case study.

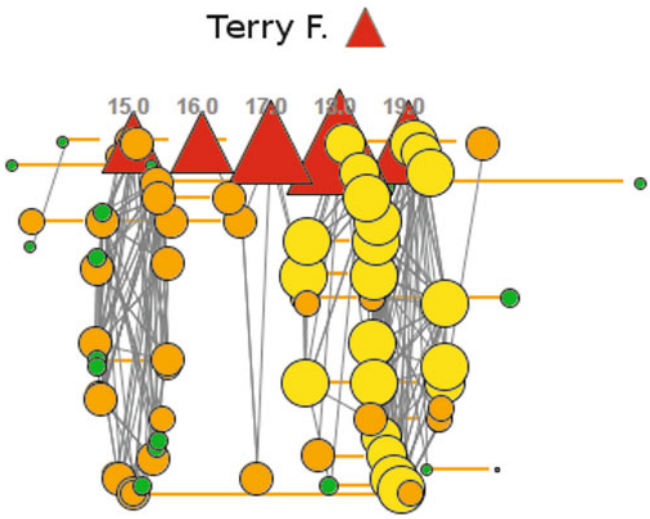


Fig. 6: Network of one specific actor and his direct contacts

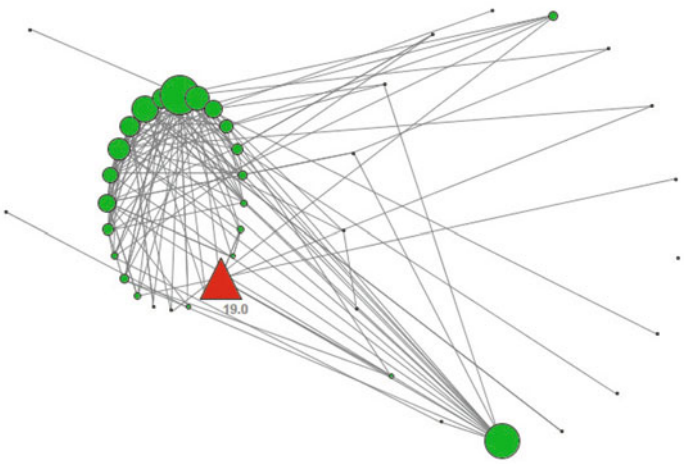


Fig. 7: k-cyclic arrangement of the actors in month 19

6 Conclusion and Perspectives

In this paper we presented a visualisation method to augment sociograms representing network communities with information about the temporal aspects and dynamics of the community. For this end we created a three-dimensional representation that can be manipulated interactively to

- Give an overview about the full actor community
- Allow a temporal snapshot of the community at a given moment
- Show the activity of any single actor in a dedicated timeline
- Integrate all the aspects in one configurable diagram.

In addition to the concept of the visualisation method, we described the data processing method, and showed our approach exemplarily with the bibliography data of the Bibtex representation of selected CSCL literature. In a more detailed case study of our recent projects we used a combination of content analysis, computation of SNA indices and the visual exploration according to this paper's approach to analyse and understand the dynamics and interaction modes within open source communities.

Besides the practical testing of the visualisation method in different contexts, such as online discussion forums, project mailing lists, and scientific publication/citation networks, we also want to use our advanced configuration features for the visual rendering via styles to highlight temporal changes of SNA traits: e.g. an actor that gains centrality over time (computed in each time slice and compared between these), should be represented with a special code, such as a bright colour showing the temporal specific of a "rising star".

From a practical point of view, we are often facing demands from industrial participants at conferences who want to use the SNA approach in their companies. The scenario addressed here starts with the problem that traditional controlling metrics fail in the case of knowledge intensive work since this type of work relies strongly on relational interaction between employees. On the other hand they have a huge amount of data as a side product of the information and communication technologies widely used in such kind of firms. In this sense one main further challenge will be the scenario where recommender systems are able to prepare complex relational data into easy understandable awareness information. Visualization is one key issue to face this challenge. Especially graph visualization enables user to capture complex information as a whole picture. Yet there is also a danger in such an approach in real life manifested in an increasing risk of misinterpretation since such visualizations may also lead to overhasty conclusions.

References

1. Austin, J. L. (1962). How to Do Things with Words, Cambridge, Mass.: Harvard University Press.
2. Batagelj, V. & Mrvar, A. (2003). Pajek - Analysis and Visualization of Large Networks. In Jünger, M. & Mutzel, P. (Eds.), Graph Drawing Software. Springer, Berlin 2003. p. 77-103
3. Gaertler, M. & Wagner, D. (2006), A Hybrid Model for Drawing Dynamic and Evolving Graphs. In Leonardi et al. (eds.), Dagstuhl Seminar Proceedings.
4. Harrer, A. (2004). Rechnergestützte Soziale Netzwerkanalyse in virtuellen Lerngemeinschaften. In: Harrer, A. & Martens, A. (Eds.), Pro-ceedings of the Workshop on Teaching and Learning Systems - The Role of Artificial Intelligence in Past, Present and Future, German Conference on Artificial Intelligence KI-2004, Ulm.
5. Harrer, A., Zeini, S., Pinkwart, N. (2006). Evaluation of communication in web-supported learning communities - an analysis with triangulation research design. In: International Journal of Web Based Communities 2006, Vol. 2, No.4 pp. 428 - 446, Inderscience.
6. Hoppe, H. U., Pinkwart, N., Oelinger, M., Zeini, S., Verdejo, F., Barros, B., Mayorga, J.I. (2005). Building Bridges within Learning Com-munities through Ontologies and "Thematic Objects". In: Proceedings of the International Conference on Computer Supported Collabora-tive Learning (CSCL2005), Taiwan, June 2005.
7. Krempel, L (2005). Visualisierung komplexer Strukturen. Grundlagen der Darstellung mehrdimensionaler Netzwerke. Frankfurt a. M.: Campus.
8. Malzahn, N., Zeini, S.& Harrer, A. (2005). Ontology Facilitated Community Navigation - Who Is Interesting for What I Am Interested in? In: Dey, A. K., Kokinov, B. N., Leake, D. B. & Turner, Roy M. (Eds.): Modeling and Using Context, 5th International and Interdisciplinary Conference, CONTEXT 2005, Paris, France, July 5-8, 2005, Proceedings. Lecture Notes in Computer Science, Volume 3554, Springer, Jul 2005, Pages 292 - 303
9. Martínez, A., Dimitriadis, Y., Gómez-Sánchez, E., Rubia-Avi, B., Jorrín-Abellán, I., Marcos, J. A. (2006). Studying participation networks in collaboration using mixed methods in three case studies. In: International Journal of Computer-Supported Collaborative Learning, vol. 1, Issue 3, march 2006. Springer.
10. de Nooy, W., Mrvar, A. & Batagelj, V. (2005). Exploratory Social Network Analysis with Pajek, Cambridge University Press.
11. Palonen, T. & Hakkarainen, K. (2000). Patterns of interaction in computer-supported learning: A social network analysis. In: Fishman, B. & O'Conner-Divelbiss, S. (Eds.), Proc. of the Fourth International Conference on the Learning Sciences. Mahwah, NJ. Lawrence Erlbaum.
12. Reffay, C., & Chanier, T. (2003). How Social Network Analysis can help to measure cohesion in Collaborative Distance Learning. In Wasson, B. Ludvigsen, S & Hoppe, H. U. (Eds.): Proceedings of the International Conference on Computer Support for Collaborative Learning. Dordrecht: Kluwer Academic Publishers, pp. 343-352.
13. Sun, L & Vassileva, J. (2006). Social Visualization Encouraging Participation in Online Communities. CRIWG 2006: 349-363
14. Trier, M. (2005). IT-supported Visualization of Knowledge Community Structures. In: Proceedings of 38th IEEE Hawaii International Conference of Systems Sciences HICCS38, Hawaii, USA, Jan 2005.
15. Wassermann, S., & Faust, K. (1994). Social Network Analysis: Methods and Application, Cambridge: University Press.

EWAS: Modeling Application for Early Detection of Terrorist Threats

Pir Abdul Rasool Qureshi, Nasrullah Memon, and Uffe Kock Wiil

Abstract This paper presents a model and system architecture for an early warning system to detect terrorist threats. The paper discusses the shortcomings of state-of-the-art systems and outlines the functional requirements that must to be met by an ideal system working in the counterterrorism domain. The concept of generation of early warnings to predict terrorist threats is presented. The model relies on data collection from open data sources, information retrieval, information extraction for preparing structured workable data sets from available unstructured data, and finally detailed investigation. The conducted investigation includes social network analysis, investigative data mining, and heuristic rules for the study of complex covert networks for terrorist threat indication. The presented model and system architecture can be used as a core framework for an early warning system.

1 Introduction

Terrorist attacks are architected after a well planned analysis. The analysis encompasses complete requirement analysis and resource allocation phases. A large volume of such valuable information is publicly available online (one estimate is 80% [13]). Intelligence agencies use many resources to collect information from various sources (Figure 1). However, a considerable amount of the valuable open source information is missed due to the limited research and development in this area [13]. Knowledge about the structure and organization of terrorist networks is important for both terrorism investigation and the development of effective strategies to prevent terrorist attacks [11].

Pir Abdul Rasool Qureshi, Nasrullah Memon, and Uffe Kock Wiil
Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute,
University of Southern Denmark
e-mail: {parq,memon,ukwiil}@mmmi.sdu.dk

However, except for network visualization, terrorist network analysis remains primarily a manual process [18]. Existing open source intelligence analysis tools do not provide advanced structural analysis techniques that allow for the extraction of network knowledge from terrorist information. In addition, open source intelligence analysis is faced with several challenges. Relevant information must be found among the enormous amount of open source information available. The collected information may be noisy: it may not be complete; it may be wrong; it needs to be filtered; etc. Once the relevant information is found and filtered, it needs to be structured, analyzed, verified, and visualized in a comprehensive manner.

It is noted that that open source intelligence analysis is not a substitute for traditional classified work. However, analysis of open source information can help to augment the analysis results available from classified information, hence providing the intelligence analysts with better support for decision making.

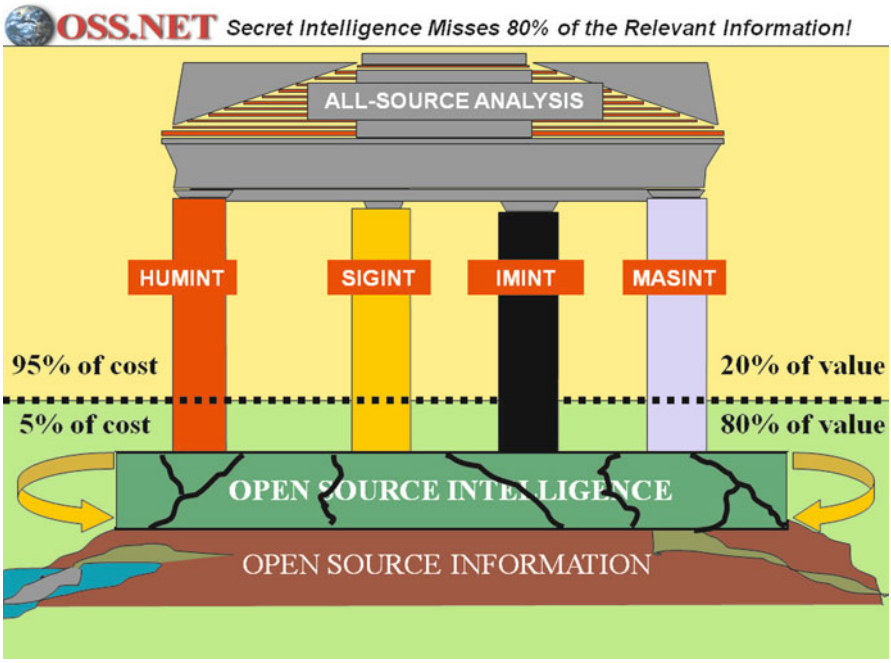


Fig. 1: The importance of open source information [13].

In this perspective, the Counterterrorism Research Lab (CTR Lab) at The Maersk Mc-Kinney Moller Institute was established. The overall objective of the CTR Lab is to specify, develop, and evaluate novel tools and techniques for open source intelligence in close collaboration with intelligence analysts.

The tool philosophy is that the intelligence analysts are in charge and tools are there to assist them [22], [24]. Thus, the purpose of the tools is to support as many of the knowledge management processes as possible to assist the intelligence analysts in performing their work more efficiently. In this context, efficient means to allow the analysts to arrive at better analysis results much faster [23]. In general, the tools fall into two categories: (1) Semi-automatic tools that need to be configured by the intelligence analysts to perform the dedicated task. (2) Manual tools that support the intelligence analysts in performing specific tasks by providing dedicated features that enhance the efficiency when performing manual intelligence analysis.

In this paper, we discuss a model for generating early warnings to predict terrorist threats and also demonstrate a limited implementation of the model. The model has two parts and a bridging application joining them. The first part uses different linguistic information acquisition and extraction modules, to structure the text into a graph. The second part, then applies various investigation techniques, such as, social network analysis, investigative data mining, statistical modeling, and heuristic rule based approaches. The bridging system between both parts takes care of the tasks like noise filtering and data source credibility, etc.

The limited implementation of the system as demonstrated in Section 10 includes a working prototype of the terrorist investigation portal which is made available to the researchers in the Counter-Terrorism Research Lab (www.ctrlab.dk), University of Southern Denmark. The terrorist investigation portal is integrated with state of art investigation system iMiner [11], which has many automated terrorist investigation and social network analysis facilities. The portal also supports manual analysis like investigating different entities within domain and analyzing their relations. The portal provides the researchers with an interface to incorporate their own knowledge about the domain dynamically in order to connect dots in a terrorist network. It serves as a platform for automatic extraction of rules and identification of user defined patterns. The terrorist investigation portal is the partial implementation of the proposed model and first necessary step towards implementation of EWAS. As the proposed model is intended to addresses all of the problems identified in Section 3 and meets the requirements discussed in Section 4, its limited implementation has already solved most of the problems and meets the major requirements and can be considered as a proof of concept.

The remainder of the paper is organized as follows. Section 2 provides an overview of state of art in the counterterrorism research approaches. Section 3 provides problems with existing systems. Section 4 discusses the functional requirements of an ideal early warning/ threat indication system; whereas, Section 5 provides an overview of the proposed system. Section 6 presents the presents the data processing phases; whereas; Section 7 discusses the detailed description of the system architecture. Section 8 presents testing and implementation strategies of early warning system; while Section 9 illustrates compliance of the system with identified functional requirements. Section

10 describes experimental results with limited functionality of the working prototype and Section 11 concludes the paper with a discussion of future extensions.

2 State-of-the-Art

Several knowledge management processes, tools, and techniques are relevant in the context of counterterrorism as shown in Figure 2 [17]. Overall, the processes in the leftmost column involve acquiring data from various sources, the processes in the middle column involve processing data into relevant information, and the processes in the rightmost column involve further analysis and interpretation of the information into useful knowledge that the intelligence analysts can use to support their decision making.

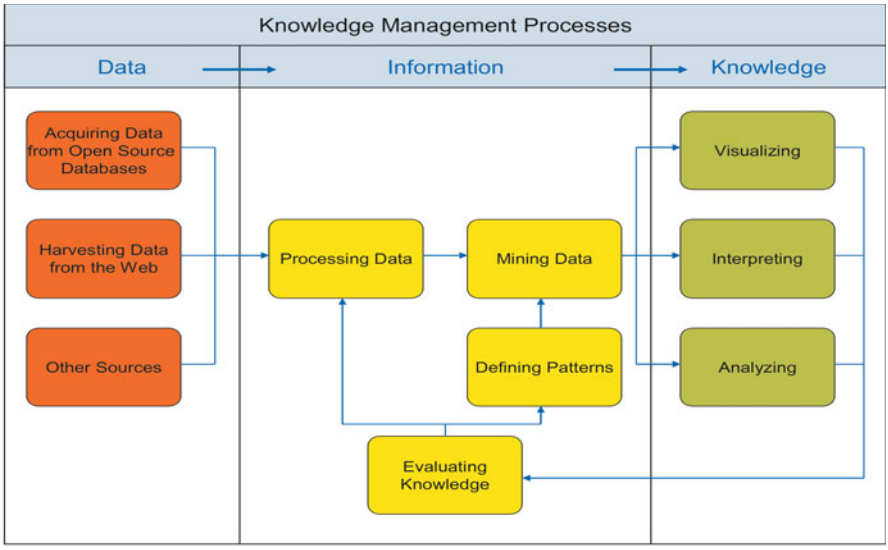


Fig. 2: Knowledge management processes for counterterrorism.

The iMiner prototype [11] includes tools for data conversion, data mining, social network analysis, visualization, and for the knowledge base [17]. iMiner incorporates several advanced mathematical models and techniques useful for counterterrorism like subgroup detection, network efficiency estimation, and destabilization strategies for terrorist networks including detection of hidden hierarchies [11]. In relation to iMiner, several collections of authenticated datasets [8] of terrorist events, that have occurred or were planned, have been harvested from open source databases (i.e., www.trackingthethreat.com).

iMiner's models and algorithms have been validated using these datasets [8] [11].

Counterterrorism research approaches can be divided into two categories: data collection and data modeling. The Dark Web Project conducted at the AI Lab, University of Arizona is a prominent example related to the data collection approach [3], [4]. The Networks and Terrorism Project conducted at the CASOS Lab, Carnegie Mellon University is a prominent example related to the data modeling approach [2]. Another modeling approach is defined in [16]. The ongoing research on EWAS combines the data collection and data modeling approaches into a holistic prototype for open source intelligence. The research involves techniques from disciplines such as data mining [5], social network analysis, [1] and [7], hypertext [14] and [6], visualization, [15] and [18] and many others. To the best of our knowledge, no other approaches provide a similar comprehensive coverage of tools and techniques to support advanced terrorist domain models. Thus, the proposed holistic research approach to support intelligence analysis is considered unique.

3 Problems with Existing Systems

Problems with state-of-the-art terrorism investigation systems are described below. Addressing these problems is necessary for effective application of investigation systems.

1. The existing systems described in the previous section are heavily dependent on the repositories of entities or keywords within the scope of the domain. These sets of entities are then used in information retrieval to identify the documents from which the information about the domain can be extracted. These repositories are seen to be managed manually; thus the definition of the domain is dependent on a manual effort. It increases the chances of ignoring the document containing entities which are not yet added to the repository.
2. Investigators working in the field have some valuable facts which are usually not used by software systems existing in industry. The obvious reason is that such type of information is not available publicly. This information which is in the mind of the investigators is very important for connecting the dots in terrorist networks. Missing such information can be costly in the analysis of terrorist networks. For instance, a terrorist might have travelled from one country to another. The investigator may know the fact from classified intelligence sources. The link of the terrorist with that country may have been missed by the existing systems working with online information because such information may not be available online. Terrorists are using the Internet in conjunction with traditional ways like travelling to someplace to meet someone, may make a phone call, may travel to another country to acquire the required skills or may go to a chemical store

to buy a chemical compound. It is obvious that such information can never be available online.
3. Existing systems are not integrated with corporate data of law enforcement agencies. The reason might be any, but without this integration a fully operational and efficient early warning generation system is difficult to achieve.
4. The existing systems do not employ social network analysis or geodesic centralities to investigate the scenario after knowledge discovery.
5. With reference to problems (2) and (3), the discussed information is usually too unstructured to be incorporated with social network analysis software. Sometimes such information is present in the form of raw files or manually drafted reports and social network analysis software expects the input to be in a comma separated value (CSV), XML or any other structured formats. That is the main reason why the information is usually missed. The text mining technologies like Gate [19] can help to structure the information but they are incapable of applying social network analysis and other heuristic approaches for disrupting the networks or digging out hidden command structure to uncover who reports to whom.

To our knowledge, no system is yet implemented to fill in the gap between social network analysis theories and text mining tools and techniques to support advanced terrorist domain models. The ideal solution which can fit in the gap should at least meet the requirements mentioned in the next section.

4 Functional Requirements

Our requirement analysis uncovered the necessity of meeting the following prerequisites for an early threat indication system:

1. The system defines a protocol to communicate and access the corporate data of law enforcement agencies.
2. The system mines the Internet for keyword specific documents.
3. Keywords can be manually added or extracted from already processed data. For example, the important attributes of known entities can be keywords for identifying new entities with same sort of attributes. The actions and specific adjectives which are within the scope of the domain and are already processed by the system can serve as the seeds for new keyword discovery algorithms.
4. The system performs advanced information and fact extraction - to convert text into structured data suitable for database storage and efficient search. Systems described in [20] and [21] have such capabilities and can be reused with special focus towards counterterrorism.
5. The system allows analysts to easily locate relationships (links) between entities (people, organizations, etc.) in unstructured text. This includes

classification of the entities (like Person, Organization, etc.) into classes (like Terrorist, Terrorist Organization, etc.) on the basis of concrete evidence, if present within the document space.
6. The system provides visualization tools to aid analysts reviewing networks of links (relationships) between entities.
7. The system performs detailed social network analysis and geodesic centrality measurements to assist analysts.
8. The system generates warnings if it encounters anything unusual, whether it is an unusual increase in number of appearances of any entity in a particular span of time or abrupt change in any of the social network characteristics. Warning generation is aided with configurable preference management to reduce the number of noisy warnings.
9. The system allows analysts to define their own rules against warn-able situations. The system generates warnings on recognition of these patterns during the knowledge discovery and investigation phases.
10. The system must provide the user with preference management facilities along with a subscription interface, so that the alerts on recognition of a specific situation, user defined pattern, or requested change in measurements of a network, is sent to the interested user via the preferred channel. Channels here can be email clients, text message receivers (mobile phones), or some electronic signals.

EWAS is an attempt to build a system that meets all of these requirements; it is also an attempt to address the problems described in the previous section.

5 The Proposed System

EWAS, as the name implies, is a system to alert intelligence agency/law enforcement personnel about any suspicious activities taking part in the world. Identifying terrorist threats require a large spectrum of data which in many cases are collected from heterogeneous sources. The process of unification, fusion, and interpretation of the collected data is crucial due to data redundancy, and specially to enable accurate predictions.

We present an early warning system which is aimed to generate early warnings against the possibility of realizing any act of terrorism to avoid such problems. The goals will be achieved by closely monitoring the information about the terrorists and the entities associated with them. Besides knowing the existing terrorists and their connections, the identification of new entities and the new connections between the entities are of premiere importance. All such information can be retrieved from a wide range of heterogeneous data sources like servers of any governmental or private organizations, web sites on Internet, news items, RSS feed files, manually drafted files, or maybe in the mind of some investigator working in the field. Thus, providing mechanisms

to acquire data from at least all of the discussed data sources is within the scope of the project. The system is built upon three main parts which are responsible for:

1. Transformation of data into workable structure.
2. Investigation of structured data to deduce results.
3. Integrating the filtered and verified structured data from the first part into the second.

For the first part, acquisition of data is very important, but does not alone suffice to complete the task. We need to transform the acquired data into some data structure, so that the complex computation can be performed and detailed investigation can take place. The suited data structure in this case is a graph. If we transform the data into graphs, then in the second part geodesic measurements and terrorist network analysis tools will help us to reach the solution. Since the data structure is a graph, the ability to visualize graphs is also an implicit requirement of the project.

In the first part, we start with searching the known keywords in the data retrieved from all of the available data sources. We locate the documents containing information relevant to our domain. Once we get our hand on data, we can semantically analyze the data to extract the information [19] of our interest from it. The structured data is then transformed into graphs of entities and relations. As extraction from all of the acquired information (semantic analysis) is computationally expensive, we filter the data on the basis of known entities to narrow down the scope of our semantic analysis operation. Although semantic analysis operation yields the entities and their relations, there can be some occurrences of the entities and relations that have low or even no importance in our context. Thus, we need to filter the data again before releasing it for further investigation and before generation of warnings in the second part.

Together with warning generation, we provide some mechanisms to help out in the manual analysis and investigation of such information, so that it can be utilized fully - by interfacing to the iMiner system [11]. This interface has facilities to perform link analysis, study analytic properties of a network, smart search any entity, and identify different patterns in terrorist networks. Investigators can formulate rules to represent these patterns to generate warnings. Thus, the raw data retrieved from heterogeneous resources is processed in different steps in order to generate warnings. We call each step as a phase. The next section discuss each of these phases.

6 Data Processing Phases

There are five major steps to process the data from heterogeneous data sources to generate warnings (if any) as shown in Figure 3. The acquisition

and extraction phases convert the unstructured data into workable structure. This workable structured data is further filtered and made available to the investigation and warning generation phases which investigate the data to generate alerts against any warn-able situation. The information generation phase is responsible for bridging the structured data from the first part (acquisition and extraction phases) to the second part (investigation and warning generation phases). Information is also filtered for unwanted relations and data from less credible data sources. The data processing phases are discussed in detail below.

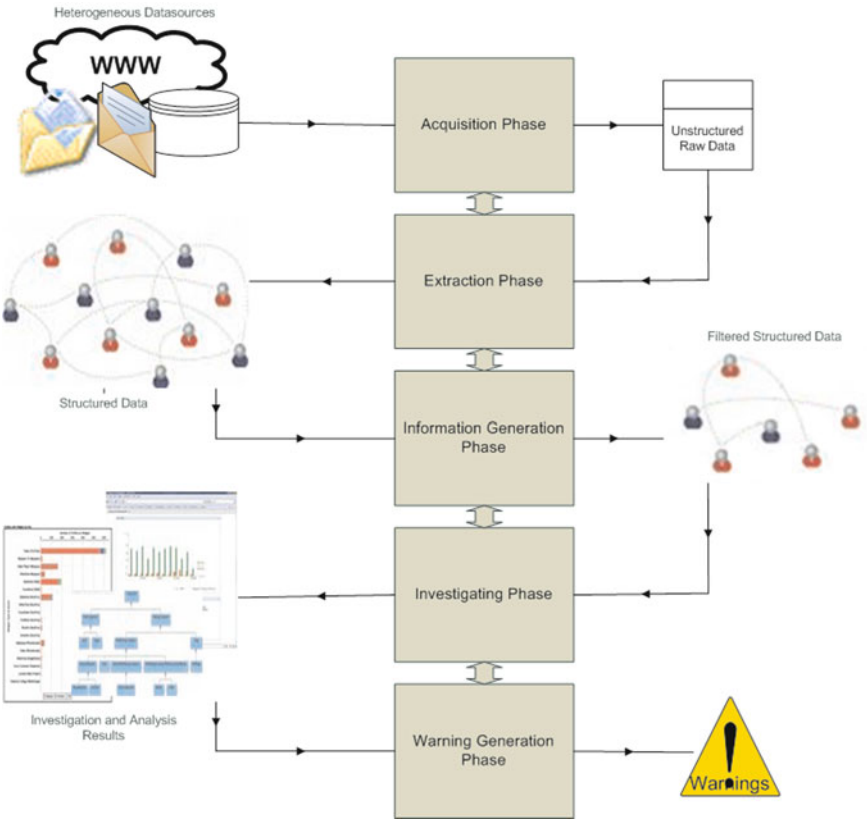


Fig. 3: Data processing from heterogeneous sources to generate warnings.

6.1 Acquisition Phase

In the acquisition phase, all of the available data from heterogeneous data sources is weighted for the presence of any of the known keywords (keywords can be entities, their relations, their attributes, etc.). The weight indicates the relativity of data in our context. If the data is worthy of further analysis, it is being short listed for semantic analysis. If the information is in a language other than English, it is being translated to English in the acquisition phase. In this phase, implementation of the processes shown in the left most column of Figure 2 takes place, where we acquire data from different data sources.

6.2 Extraction Phase

Data that has been short listed is analyzed semantically to extract the information hidden in the data. The extracted entities and their relations are kept in the data store. As all of the entities and their relations are identified during the extraction process, and since the information may contain some unwanted entities or relations, the data store in which this information is kept is called a dirty database. It is here that we transform the information within data into a graph data structure.

This phase, together with the information generation phase, implement the processes shown in middle column of Figure 2. In this implementation, thorough semantic analysis of data, identification of patterns which can be used to transform data into useful information and filtering of unwanted data, takes place.

6.3 Information Generation Phase

In this phase, the information in the dirty database is further filtered and only those entities and relations which are related to our domain are selected and then made available for further investigation by the Publisher Application. Different attributes of data like credibility of data source, number of appearances of same data, etc, are also considered by the Publisher Application (see Figure 4) which can run in automatic or semi automatic mode.

6.4 Investigating Phase

In this phase, different investigations are carried out. The information is evaluated with the terrorist network analysis techniques like dependence central-

ity, position role index, and also geodesic measurements, like average path length, clustering coefficient, and density. Role analysis detecting different roles assigned to different nodes of a graph/ network/ cell, and identifying command structures are some examples of investigations that are carried out in this phase. The measurements to estimate the effectiveness of different destabilization strategies for particular terrorist network can be made. In this step, we track and monitor changes in the different characteristics of a terrorist network over a particular span of time. Also the same data is made available to investigators for manual investigation. The investigation system also exposes interfaces to integrate with state-of-the-art investigation frameworks like iMiner [11] to carry out the further investigation in this phase.

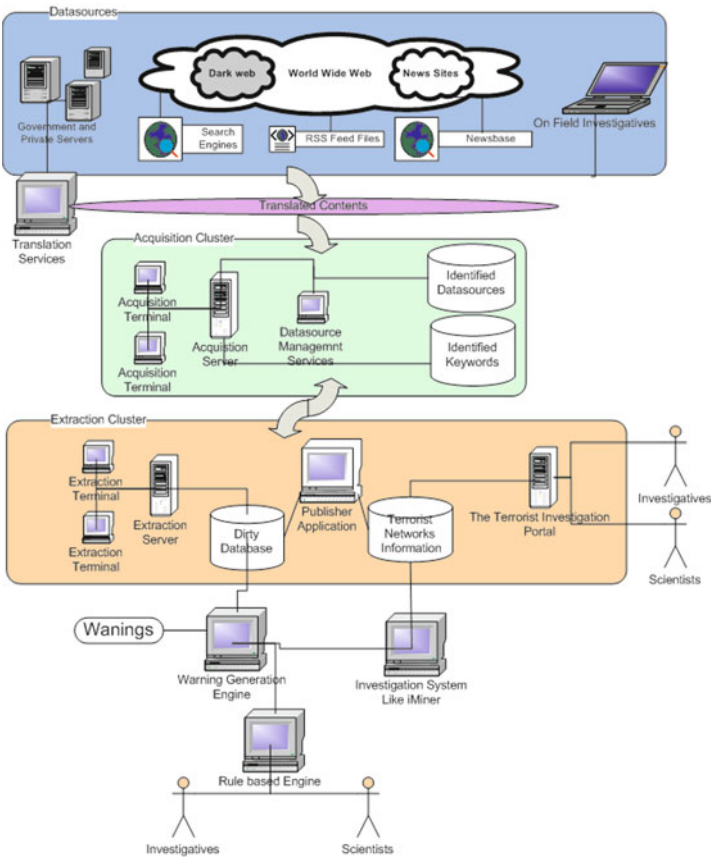


Fig. 4: EWAS system architecture.

6.5 Warning Generation Phase

If the warning generation engine encounters any abrupt shift in the measurements and characteristics of the graph, and the extent of shift is high as to a dangerous level, the warnings are generated. The warning engine also identifies the presence of user defined patterns within graphs with the help of accompanying rule execution engine to generate warnings. The users are provided with access to the warning generation engine, to input and test their researched patterns and theories, all of which are kept with the warning generation engine as rules and are used in the warning generation phase.

The components which are responsible for the transformation of data from one phase to another are described in the next section along with the overall system architecture.

7 EWAS System Architecture

The EWAS system is actually a system of subsystems, all working together to assist the warning generation engine to do its role, i.e., generate warnings. Each of these subsystems (as shown in Figure 4) includes complex network applications running in clustered environment with the standards of parallel processing to simple web applications for user interfacing. That is why we used the word cluster for these systems. A description of each of these subsystems is given below.

7.1 Acquisition Cluster

The acquisition cluster is responsible of acquiring information from the Internet as well as many government and private databases available online. It is actually a system of wrappers and interfaces corresponding to the data sources. These wrappers can be extended to support any type of data source, if necessary. The acquisition application uses the search engines to search the web for information for any of the required entities and may contain the specialized routines to carry out dark web analysis. The time oriented treatment of data sources like news web sites and Really Simple Syndication (RSS) feeds is within the scope of acquisition system. Therefore, it is equipped with smart readers which track the dates and times related to different extractions made. Information in languages other than English is translated into English before actual acquisition takes place. This is done by requesting translations from available third party translation services like Google translate. The acquisition system short lists the related information for further semantic analysis on the basis of known keywords kept in the keywords database.

The acquisition system is a parallel computing system complying with plug-n-play architecture. We can extend the number of acquisition terminals in real time depending on processing needs. The acquisition system will interact with the data source management system, the other system running in the acquisition cluster to identify the data sources, from where data is needed to be acquired.

The data source management system is responsible for maintaining the list of the registered data sources, their access details, the priority list with respect to preference for a data source, and credibility. It also records the different experiences about the data sources like frequency of updating and the structural information about that data source (which part of the page contains the real information) necessary to wisely schedule and smartly control the acquisition system for acquiring information from that particular data source.

7.2 Extraction Cluster

The extraction application performs semantic analysis of the short listed text pages. The extraction application is parallel computing system. The number of terminals may be extended to feed the processing needs. The extraction application processes the acquired data, identifies the entities and their relations, and transforms it into the graph data structure. Services of many linguistic models programmed and shipped with Gate [19] are integrated into the extraction application. It also updates keywords after the identification of new entities or their relations into the keywords database which is being used by the acquisition system. The extraction application saves all of the extracted information, even information containing unrelated entities in its database, which is a so called dirty database. The information from the dirty database is published into a final database containing terrorist information which is being used by the Terrorist Investigation Portal (TIP); an interface to expose the data for manual analysis and investigation software like iMiner [11].

The publisher application is the other application inside the extraction cluster. This application runs in any of two modes, automatic or manual. In automatic mode, it evaluates the information in the dirty database with a set of rules. Information that complies with the rules will be published automatically. These rules represent different heuristic approaches to verify and validate the information. The "relation which is reported by five different data sources" and the "entity reported by New York Times data source" are examples of the data which is validated automatically with the help of these rules in automatic mode. In manual mode, it shows the information to the end user, who can be any domain expert and lets the user decide whether to pub-

lish the entities or not. The information filtered by the publisher application is kept in our production database, called terrorist networks database.

The information in the terrorist networks database is made readily available to all manual investigators and scientists/researchers and a variety of systems by TIP, which is composed of a set of data services and a portal application. These services use investigation systems to carry out geodesic investigation and terrorist network analysis. The TIP application is aimed to be an online system open to all registered users. At present it is just published at University of Southern Denmark's internal domain for testing purposes. Presently, it has functionalities of not only visualizing the graphs formed and updated automatically on daily basis, but also to carry out different social network analysis techniques like detection of hidden hierarchy. It also provides users with the interface that allows them to input their investigation patterns and to test their theories on terrorist networks.

7.3 Investigation System

The TIP exposes the data services, which enable the investigation system iMiner [11] to investigate the terrorist networks with proven technologies. The social network analysis encompasses the techniques determining the change in dependence of a network on a particular node, role analysis by evaluating the nodes of the network with respect to position role index, identifying the change in role with the evolution of a network, measuring values of geodesic centralities (like degree, eigenvector, betweenness, closeness, etc.) calculating the average path length, clustering coefficient, density, efficiencies, etc. of a network [11]. All this information can be used to define rules which that are used to trigger warnings if any network has evolved to the defined state.

7.4 Warning Generation System

The warning generation engine with the help of the investigation system monitors the changes in the geodesic measurements. The changes in any of the measurement may cause the warning engine to generate a warning. The system is equipped with a rule execution engine. The rule execution engine can be configured with rules representing threshold values and user defined patterns. The successful execution of the rules governing warning generation controls the number of noisy warnings. Rules can also be configured to reflect the changes in relationships of entities or addition of a node to a graph like any other social networking sites. All of these warnings are sent to users subscribed for such patterns or interested in such rules. The warnings are sent to users on their preferred devices like an email client or mobile phones (SMS

receiver). The rules representing any sort of warn-able situation described above are named as Warning Generation Rules (WGRules). The anatomy of WGRules is described below.

7.5 Warning Generation Rule Anatomy

As mentioned, the rule execution engine embedded into the warning generation engine is capable of processing rules submitted by analysts. These rules provide a channel for analysts and scientists to input and experiment their theories about the evolution of terrorist networks or different patterns of warn-able situations. The users may reuse any of these rules by specifying it as precondition for a new rule. This technique provides the mechanism of rule chaining, i.e., connecting the output of one rule as input to another rule. Rule chaining makes it easy for the user to transform his/her thoughts to the system. Figure 5 shows the anatomy of a typical WGRule.

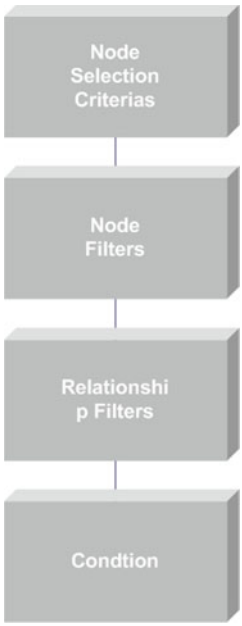


Fig. 5: Warning generation rule anatomy.

Node Selection Criteria. This part specifies the criterion which should be matched with the context of any node. For example, if the analyst wants to carry out the analysis for any particular bank robbery, he defines it in the node selection criteria. By such node selection criteria, the warning generation

system limits the scope execution of this rule to only bank robbery cases. The patterns related to bank robberies are loaded into the system while executing this rule. The set of preconditions about the successful execution of different rules can also be specified in this part of WGRule. So if any of the rules specified as precondition for a particular rule fails, the system does not execute that particular rule. Even no context is established for execution of such rule.

Node Filters. In this part, the analyst defines the nodes which constitute the network over which the rule will be executed. This part of WGRule can be used to specify filtering logic for the nodes within the context. Nodes can be filtered on the basis of different attributes, associated data sources, or different positions in a network (like at distance 5 from a particular node).

Relationship Filters. This part shortlists the type of links and relationships to be included in the network over which the rule is going to be executed. Relationship filters help in situations when a rule is needed to be executed on one dimensional information about any entity like educational information (the network containing connections of any entity with universities and educational institutes) or travel information (the network containing connections of any entity with different countries or cities visited). Once the nodes and relationships filters are applied, a true network is formed comprising of selected nodes and relations within the established context. Now this network can be evaluated in light of social network analysis and geodesic measurements.

Condition. Condition is the part of the rule in which the user identifies the calculations and threshold values that should be considered for successful execution of rule. For example "If a node with at least four nodes reporting to it [10] is captured and network evolution is observed in the European region, the rule will be executed true" can serve as a valid condition for WGRule. Likewise the conditions can be based on different social network measurements or attribute and role analysis, as well as other calculations which are necessary to indicate the warn-able situation.

WGRules represent user defined patterns which are input by users and are evaluated by the rule execution engine for generation of warning. The testing of these rules require huge amount of structured data about terrorist networks operating in the present world. So we have partially implemented the system with special emphasis on data collection, verification and validation part. This data is published and served to our investigation systems through terrorist investigation portal. The testing and implementing strategy for the EWAS system to be incorporated in real-time environment of Law Enforcement Agencies is described in the next section.

8 Testing and Implementation Strategy for EWAS

Presently, the system is under developed and parts of it are ready for testing. The main parts of the acquisition, extraction, and investigation systems have been implemented. The publisher application is functional and the Terrorist Investigation Portal has been made available to CTR Lab members on the internal network at University of Southern Denmark. At present, information collected from open source intelligence sources has been processed and published in the Terrorist Networks Database. The information is made available to the state-of-the-art investigation system (iMiner [11]). The facilities to conduct manual analysis are provided to researchers working in CTR Lab via the Terrorist Investigation Portal and iMiner. The idea (as presented in Figure 6) is to test the validity of the system on the data acquired from open sources; and at the implementation time the system will be integrated with corporate data of the law enforcement agencies. After implementation, the system will thus incorporate the fusion of open source information and corporate databases of law enforcement agencies. As the system will reside within the boundary of law enforcement agencies, the different data access limitations can be relieved. The compliance of the system with requirements (Section 4) and the solutions of the problems (Section 3) are evaluated in the next section.

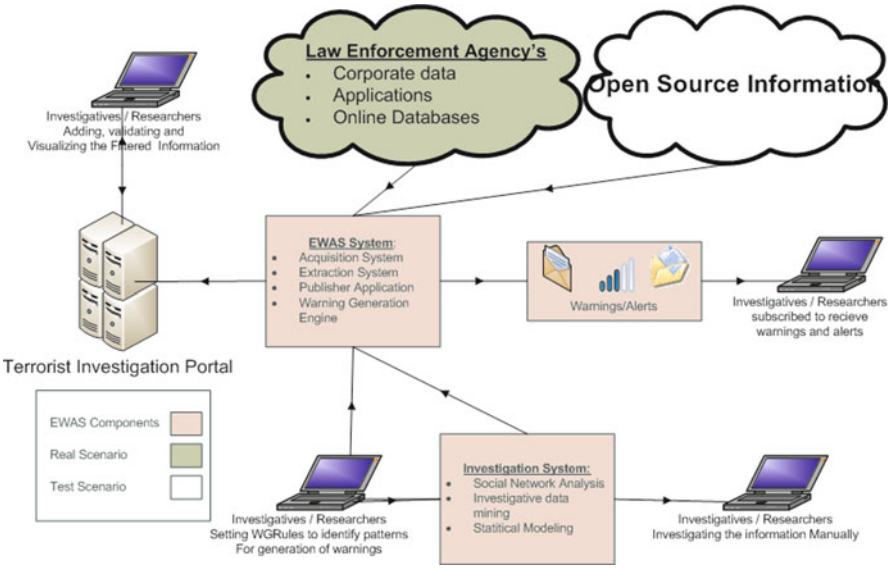


Fig. 6: EWAS system architecture.

9 Compliance with Requirements

The system in the implementation phase will carry the benefits of open source information and achieve its integration with corporate data of investigative agencies. The change in idea (i.e., to test and train the system over open source information and provide it to intelligence rather than expecting access of corporate data from intelligence agencies) has already solved major problem. The 80% of valuable information which agencies miss [13] is made available to them to support their analysis. The automated information extraction combined with sophisticated social network analysis, incorporating data validation, verification, attribute analysis of entities and automatic addition of keywords or seeds has solved all of the major problems. The system even in test scenario supports the manual analysis, graph visualization, and knowledge extension by researchers. So the system meets the requirements described in Section 4. However, the system has not achieved optimum performance in the warning generation part. It is due to the absence of training materials in the counterterrorism domain to optimize the advanced linguistic and investigation models programmed in EWAS. We believe the optimization at this part can be achieved by the time with the increase of information in our database. The next section documents the preliminary results achieved by the EWAS system and its interface exposed by the Terrorist Investigation Portal.

10 Experimental Results

Presently we have a working prototype of EWAS which is in the testing phase. It is capable of acquiring data from open sources[1] transforming it into graph structures, analyzing links between entities; calculate the geodesic and various centralities of a network. Its interface (Terrorist Investigation Portal) is available for network visualization and manual analysis within the internal network of University of Southern Denmark.

Figure 7 is an example of a thwarted terrorist plot known as Bonjika. We harvested (as described in [8]) the information about the Bojinka plot from open source information, used iMiner's investigation modules aided with manual analysis by researchers through Terrorist Investigation Portal. The automatically detected command structure by the modules of iMiner from the open source information about the plot is shown in Figure 8, where it is crystal clear that Khalid Shaikh Muhammad is the leading member of the large cell of terrorists.

[1] www.trackingthethreat.com
and www.globalsecurity.org

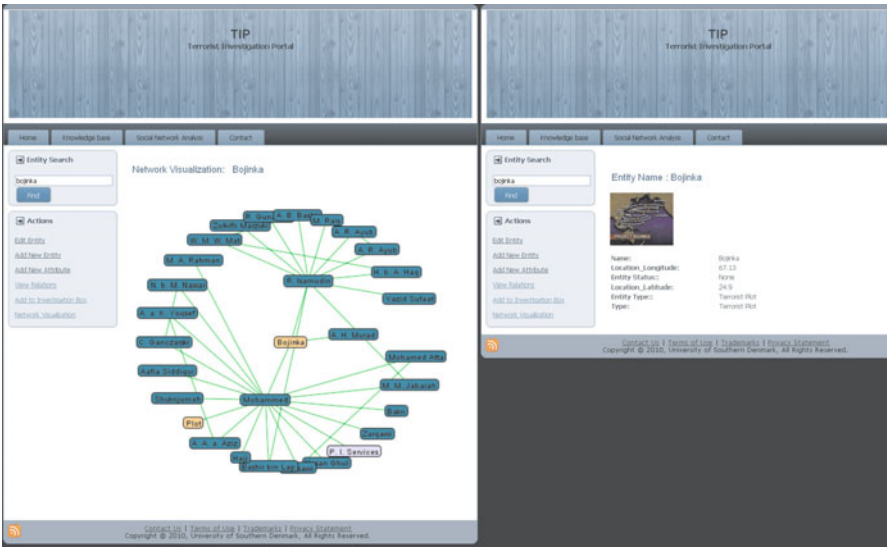


Fig. 7: The Terrorist Investigation Portal.

11 Conclusion and Future Extensions

In this paper we discussed various state-of-the-art systems in the field of terrorist investigation and their shortcomings. We highlighted the shortcomings of the systems to integrate with corporate databases of law enforcement agencies and to accommodate the manual analysis carried out by investigative agencies. The state-of-the-art systems lack in including 80% [13] of the available information in the investigation.

In this paper, we analyzed the state-of-art and discussed various problems associated with it. We identified functional requirements which an ideal terrorist investigation system must meet. We presented a model addresses the identified problems with state-of-art and meets most of the identified requirements. The model comprises of three parts. First part transforms the data into workable datasets and second part to carry out the detailed investigation. The third part filters and integrates the data from the first part to the second part. The transformation of data from unstructured data to structured workable data sets is accomplished by information retrieval and extraction systems. The structured information is published by publisher application which takes into account different attributes of information like data source credibility and noise filtering, etc., before publishing it. The second part investigates the published data using the techniques from various state-of-the-art investigation systems like iMiner [11].

We also described in this paper various phases in which data is processed in our model to conduct investigation and generate warnings.The warning gen-

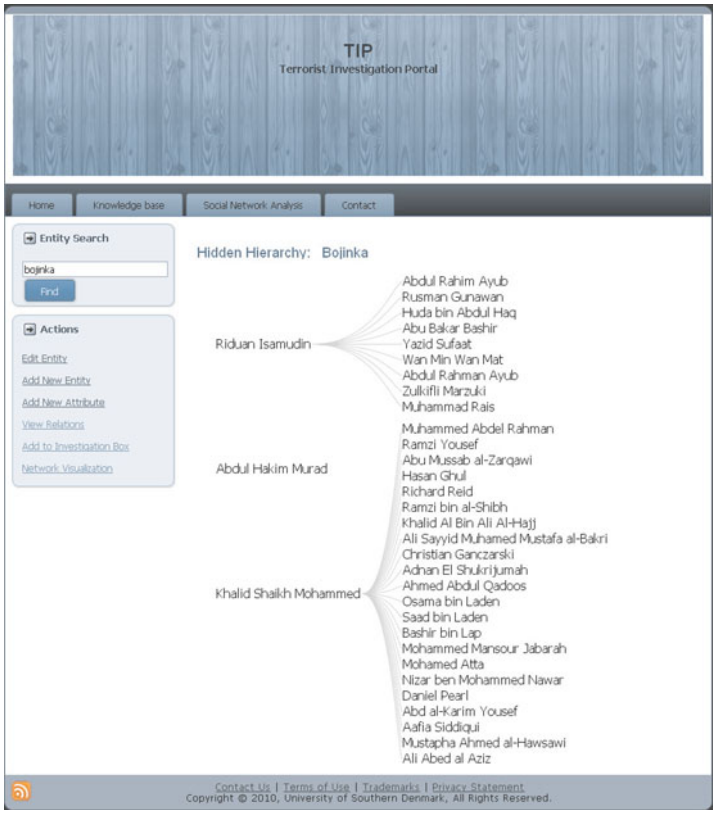


Fig. 8: Hierarchical chart of Bojinka Plot.

eration is accomplished with the help of a rule based engine, which executes WGRules and generates warnings over changing characteristics of graphs. The rules can encapsulate different user defined patterns which the investigating person may set according to his/her preferences.

We introduced a working prototype of Terrorist Investigation Portal; a partial implementation of the proposed model. We conducted the experiment to collect the information from open source information related to a past terrorism plot named Bojinka. The data collected was analyzed with the help of terrorist investigation models implemented in state-of-the-art investigation system iMiner [11]. We identified the hidden command structure in the network of the terrorists involved in the attack. The results we achieved through investigation are in excellent agreement with reality as Khalid Shaikh Muhammad is leading a major cluster of the graph. The experiment with partial implementation showed that the approach of combining information extraction modules with investigation systems can save a lot of time and add an additional benefit of incorporating 80% [13] of the data, which is not usu-

ally used. Thus, the partial implementation can serve as a proof of concept for the proposed model which combines open source intelligence with social network analysis applications.

In the future, the analysis carried out on the terrorist investigation portal, will be transformed to form training data. This training data will be used by automatic rule execution and pattern identification module for learning. The sufficient amount of training is necessary to optimize the generation of early warnings. Once the sufficient quantity of analysis is gathered from terrorist investigation portal, we will incorporate it to implement the proposed model.

References

1. Carpenter, M. A., and Stajkovic, A. D. 2006. Social Network Theory and Methods as Tools for Helping Business Confront Global Terrorism: Capturing the Case and Contingencies Presented by Dark Social Networks. In G. Suder (Ed), Corporate Strategies Under International Terrorism and Adversity, 7-19. UK: Edward Elgar Publishing.
2. Carley K.M. et al. 2006. Toward an Interoperable Dynamic Analysis Toolkit. Decision Support Systems 43(4): 1324-1347.
3. Chen H. et al. 2008, Terrorism informatics: Knowledge Management and Data mining for Homeland Security, Springer.
4. Chen H. et al. 2008. IEDs in the Dark Web: Genre Classification of Improvised Explosive Device Web Pages. In proc. IEEE ISI 2008.
5. Devlin, K., and Lorden, G. 2007. The Numbers Behind NUMB3RS: Solving Crime with Mathematics. Plume.
6. Engelbart, D. C. 1962. Augmenting Human Intellect: A Conceptual Framework, Summary Report. AFOSR-3233, Stanford Research Institute.
7. Gloor, P. A., and Zhao, Y. 2006. Analyzing Actors and Their Discussion Topics by Semantic Social Network Analysis. Information Visualization. IV 2006, pp. 130-135.
8. Memon, N., Wiil, U. K., Alhajj, R., Atzenbeck, C., and Harkiolakis, N. 2010. Harvesting Covert Networks: The Case Study of the iMiner Database. Accepted for the International Journal of Networking and Virtual Organizations (IJNVO). InderScience (to appear).
9. Memon N.; Larsen H.L.; Hicks D.; Harkiolakis N. 2008. Detecting Hidden Hierarchy in Terrorist Networks: Some Case Studies: Lecture Notes in Computer Science, vol. 5075/2008, pp. 477-489.
10. Memon N.; Hicks D.; Larsen H. L.; Uqaili M.A. 2007. Understanding the Structure of Terrorist Networks, In International Journal of Business Intelligence and Data Mining, vol. 2(4), pp. 401-425.
11. Memon N. 2007 "Investigative Data Mining: Analyzing, Visualizing and Destabilizing Terrorist Networks", PhD dissertation, Aalborg University.
12. Memon N.; Hicks D.; Larsen H. L. 2006. How Investigative Data Mining Can Help Intelligence Agencies to Discover Dependence of Nodes in Terrorist Networks. Lecture Notes in Computer Science, vol. 4632/2006, pp. 430-441.
13. Steele, R. D. 2006. The Failure of 20th Century Intelligence. c
14. Shipman, F. M., Hsieh, H, Maloor, P., and Moore, J. M. 2001. The Visual Knowledge Builder: A Second Generation Spatial Hypertext, In Proc. of the ACM Hypertext Conference, pp. 113-122. ACM Press.
15. Thomas, J., and Cook, K. 2006. A Visual Analytics Agenda. IEEE Computer Graphics and Applications 26(1), 10-13.

16. Tsvetovat M., and Carley K. M. 2005. Structural Knowledge and Success of Anti-terrorist Activity: The Downside of Structural Equivalence. Journal of social structures 6(2).
17. Wiil U.K., Memon, N., Gniadek J. 2009. Knowledge Management Processes, Tools and Techniques for Counterterrorism. International Conference on Knowledge Management and Information Sharing (KMIS 2009), (Funchal, Portugal, October). INSTICC Press, pp. 29-36.
18. Xu J., Chen H. (2005) "CrimeNet Explorer: A framework for criminal network knowledge discovery ", ACM Transactions on Information Systems, Vol. 23(2), pp. 201-226.
19. Damljanovic D., Agatonovic M., Cunningham H. 2010.: Natural Language Interfaces to Ontologies: Combining Syntactic Analysis and Ontology-based Lookup through the User Interaction. In Proceedings of the 7th Extended Semantic Web Conference (ESWC 2010), Springer Verlag, Heraklion, Greece, May 31-June 3, 2010.
20. Maynard D., Funk A.,Peters W. 2009. Using Lexico-Syntactic Ontology Design Patterns for ontology creation and population. In Proceedings of ISWC Workshop on Ontology Patterns (WOP 2009), Washington, 2009.
21. Ning Li, Jun Rao, Eugene Shekita. 2009. Leveraging a scalable Row Store to Build a Distributed Text Index. In Proceedings of the first international workshop on Cloud data management (CloudDB 09) Hong Kong, China, pp. 29-36.
22. Qureshi P.A.R., Memon N., Wiil U. K. 2010. EWaS: Novel Approach for Generating Early Warnings to Prevent Terrorist Attacks in the Proceedings of 2010 Second International Conference on Computer Engineering and Application (ICCEA 2010), Bali Island, Indonesia, IEEE Computer Society, pp. 410-414.
23. Qureshi P. A. R., Wiil U. K., Memon N. 2009. Modeling Early Warning System to Predict Terrorist Threats: Preliminary Results. In the Proceedings of the 10th International Symposium on Knowledge and System Sciences, Hong Kong, China, 2009. pp. 179-190.
24. Memon N., Wiil U.K., Qureshi P. A. R. 2009. Design and Development of an Early Warning System to Prevent Terrorist Attacks. In the Proceedings of International Conference on Artificial Intelligence and Neural Networks, Beijing, China, 2009. IEEE Press, pp. 222-226

Complex Dynamics in Information Sharing Networks

Bruce Cronin

Abstract This study examines the roll-out of an electronic knowledge base in a medium-sized professional services firm over a six year period. The efficiency of such implementation is a key business problem in IT systems of this type. Data from usage logs provides the basis for analysis of the dynamic evolution of social networks around the depository during this time. The adoption pattern follows an "s-curve" and usage exhibits something of a power law distribution, both attributable to network effects, and network position is associated with organisational performance on a number of indicators. But periodicity in usage is evident and the usage distribution displays an exponential cut-off. Further analysis provides some evidence of mathematical complexity in the periodicity. Some implications of complex patterns in social network data for research and management are discussed. The study provides a case study demonstrating the utility of the broad methodological approach.

1 Introduction

Developments in information technologies have increased efficiency in much organisational communication. This is not simply at the level of dyadic information exchange, by email, instant messaging and the like, but more significantly at systemic levels in the storage and retrieval of vast quantities of information central to core transactions. But these efficiency gains have been largely concentrated in those information processes that are most readily codifiable, explicit and thus relatively low value. Information technologies have made notably less impact on systemic communication processes at the

Bruce Cronin,
University of Greenwich Business School, Park Row, London SE10 9LS, UK
e-mail: B.cronin@greenwich.ac.uk

centre of higher value knowledge sharing activities where human interaction is necessarily more complex [1-3]. Snowden [3] argues that this is a symptom of an over-emphasis on technology as the critical link between individual knowledge rather than the interpersonal relationships themselves.

This study accordingly explores the evolution of interpersonal relationships during the roll-out of an electronic knowledge base in a medium-sized professional services firm, drawing from usage logs. It employs a metaphor of "social complexity" to make sense of the dynamic pattern and utilises a range of techniques, including social network analysis, to pinpoint critical aspects of the evolving relationships. The key aim is to identify the critical conditions that initially inhibited and then led to the widespread usage of the knowledge base, a key business problem in the field.

A combination of methods is used to this end, including trend analysis, Fourier analysis, social network analysis and regression analysis of the latter two. While Fourier analysis is typically used to identify trends in large continuous datasets as in this study [4] and social network analysis has been increasingly applied to information problems, combinations of the two are rare [5]. This study offers a methodological innovation in regressing data from the two approaches to identify critical change conditions. It provides a case study demonstrating the utility of the broad methodological approach.

2 Literature

The challenge of reaping efficiency gains from information technology in high-value systematic processes is rooted in the complicated nature of social interaction. In an attempt to unpack the various dimensions of this problematic interaction, Monge and Contractor [6] identify a range of theoretical motivations that may condition human engagement in systemic communication processes. Self-interest models include the accumulation of social capital [7] and the minimisation of transaction costs [8]. Public goods models emphasise the mutual interest in contributing knowledge resources to depositories when others do [9]. Social exchange theory provides a model for the growth of interaction, when contributions are reciprocated [10], to which could be added resource dependency and uncertainty reduction models [11]. Homophily and proximity models focus on the way shared attributes condition interactions [12]. Contagion models emphasise the way the structure of interaction itself facilitates or hinders isomorphic behaviour [13, 14].

For Monge and Contractor [6], the multiplicity of motivations helps explain why the use of electronic depositories has fallen well short of expectations. They argue that electronic information depositories primarily serve to publicise the existence of expertise, which is then pursued by other means [see also 15]. At the same time, contributions to such depositories are also variously and contradictorily motivated. Preliminary empirical investigation suggested

public goods motivations were predominant in contributions, with rewards to generalised participation coming from greater knowledge differentiation and awareness of sources of expertise [16]. However, subsequent contagion models continue to be salient [17].

If participation in systemic communication, such as engagement with corporate information depositories, is a contagious process, then a staged evolution might be expected, such as the innovation diffusion model [18]. Gongla and Rizzuto [19], for example, suggest a five stage process from potential, to building, engagement, activity and adaptation (see Table 1). Here, human interaction is presented as an unproblematic component but even the innovation diffusion model posits critical roles for different groups of people at different stages, particularly innovators and early adopters. Extending this recognition, Afuah [20] suggests five key roles in the innovation process:

- Idea generators, who synthesise a variety of information to generate new marketable products, employing disciplinary expertise and broad cross-disciplinary scan [21, 22];
- Gatekeepers (interfirm) and Boundary Spanners (intrafirm), who link internal resources and processes with external information [23, 24];
- Champions, who evangelically adopt and promote an innovation at some risk [21, 25, 26];
- Sponsors, coaches or mentors, who provide quiet senior-level support, access and protection [21, 27]; and
- Project managers, who provide methodical specification of the tasks needed to implement the innovation.

But a process of human interaction involves much more than the specification of key roles or tasks; the nature of the interaction is critical. Isherwood [28], for example, found take-up of collaborative knowledge bases driven by top-down leadership where it met real business needs and where there was thorough training. Similarly, Ayers [29] found collaboration for quality improvement dependent on "cultivating trust, attendance to the human dimension, nonlinear development, attendance to organizational culture, integrated philosophy of quality improvement, and a focus on process and outcome measurement to drive change" (p. 234). Yet, in light of Monge and Contractor's [6] discussion of the multiplicity of motivations in a social setting, can adoption really be expected to simply follow from a supportive organisational culture and a clear focus on the business case? The persistent difficulties in managing this activity suggest a much more complex situation.

Accordingly, there would seem to be value in exploring the metaphor of social complexity that has recently entered some areas of management research. The metaphor draws from the study of complex systems in the mathematical and natural sciences, those between discoverable order and randomness. In such systems, processes are not readily predicted, small changes can have large effects yet broad regularities emerge from the small micro-interactions but are not reducible to these. The regularities are non-linear, relatively stable

Table 1: A model of staged evolution of IT-mediated knowledge sharing.

Stage	Potential	Building	Engaged	Active	Adaptive
Definition	Formation	Self-definition and formalised operating principles	Process improvement	Understanding and results from collective work	Use of knowledge for competitive advantage
Functions	Connection	Memory and context	Access and learning	Collaboration	Innovation
Behaviour	People find one another and link	Share experiences knowledge, create roles norms	Commitment, outreach, promotion of knowledge-sharing	Problem-oriented workgroups, connections with other groups	Work together to advance knowledge in their field. Sponsor new communities
IT Role	On-line directories forums	Repository, classification, collaborative environment	Portals, yellow pages, surveys	Team rooms, issue discussion boards, integration with corporate software	Pilot new technology, Technology transfer, Integration with external software

Source: Adapted from Gongla and Rizzuto [19].

patterns that the system tends to settle around, "attractors" [30]. [1] Complex systems are broadly characterised by the presence of many densely interconnected parts and the dependence of systemic outcomes on widely distributed activity among the parts. Social complexity remains a metaphor, however, firstly because human activity is not reducible to the mathematical properties of a closed system; social systems are open. Secondly, social interactions are invariably multidimensional and vary with interpretation and intentionality [31].

Participation in systemic communication processes would seem to be characterised by many of these features. The variety of contradictory motivations discussed by Monge and Contractor [6] seem to limit the possibility of explanation in terms of simple linear relationships. The contagion processes evident in the adoption of this practice highlight the interconnectedness of

[1] An attractor is a region of systemic activity rather than a single point in the system. An analogy of a 'point' attractor is a marble in a fruit-bowl that may spin around the bowl in an unpredictable way but settles on a broadly predictable point in the bowl. An analogy of a "complex" attractor is a rabbit in a cage worried by a barking dog outside. The activity of the rabbit and the dog is broadly predictable but the specific movements of rabbit or dog are not. Yet a pattern of behaviour emerges from these conditions. In both cases, the attractor is not a single point, as a magnet may attract iron filings, but comprises the specific conditions: in the first case the marble, the bowl and gravity; in the second case, the combination of the rabbit, the dog and the cage [30].

participants, while the identification of a range of different key actors at various stages points to the systemic dependence on widespread participation and non-linear phase changes. So, identification of emergent, systemic patterns of behaviour and the conditions underpinning these are likely to provide a fruitful line of enquiry for the understanding of these processes.

3 Methods and Data

To explore the applicability of the social complexity metaphor to the development of high value systemic communication processes, a case study was undertaken of the introduction of an electronic information depository in a medium-sized professional services firm over a six year period. The adoption of this practice innovation was motivated largely by a desire to minimise paper-based corporate filing systems and a general aspiration to enhance information flows and knowledge sharing. While the introduction of this system fell well short of the standards of good practice in information systems, with little formal system specification or user needs analysis, this is arguably a reasonably typical case of the introduction of such systems in such organisations.

The data for the study comprised the usage logs from the knowledgebase from establishment in 2002 through the calendar years to 2008. The 108,000 transactions logged from 232 authorised (potential) users represents contributions or searches for information in the organisation that cannot be found more readily by other means.

After reviewing descriptive statistics and considering the form of usage trends, complex dynamics in the data were analysed. First, a Fourier analysis was undertaken to model any discontinuities found. Fourier observed that any curve can be represented as a sum of sine and cosine terms; a Fourier analysis measures the frequency of oscillations in the sequence of interactions, the sine representing the amplitude of each oscillation [32]. Deconstructing the complex usage/user function into sine waves then provides the means of identifying the root of the function or attractor underpinning the relationship. The most frequent amplitude in the spectrum then points to the root of the function, or the complex attractor in this conception.[2] The users associated with this mode amplitude in their pattern of usage were then identified.

A difficulty with the Fourier analysis is selecting the appropriate period for analysis. While the entire set of transactions arguably represents a single complex system of interaction, it is more likely that a series of complex situations have succeeded each other, each with distinct attractors. The conditions and pattern of usage at the introduction of the knowledgebase are likely to be

[2] The analysis determines the formula for the curve at each point in the series of data points in the form a + bi. The absolute value of this complex number is derived and then the modal range of these determined with a histogram.

quite different than those after widespread adoption. However, the attractors are likely to operate over a prolonged period as it takes time for posters to deposit relevant information and for new users to discover the information available and for existing users to contribute or discover additions, although this will be a function of the number and extent of usage rather than a linear time function. Preempting the results, the daily pattern of usage suggests something of a quarterly and annual periodicity. As the analysis requires sample sizes that are an exponent of two, alternative models were thus tested based on series of 64, 128 and 256 days.

Following the Fourier analysis, a social network analysis was then undertaken at six distinct points during the period to attempt to identify key figures or roles in the network of users' postings and readings and to assess the extent of information sharing between organisational divisions. The clustering coefficient was also calculated for evidence of power-law distributions [33-40].[3] The latter was determined in three steps. First, the relational nearness (shortest path) of each node to each other in the network was derived. Then the average nearness of all nodes in one division to all nodes in each other division was calculated and the logarithms of these plotted.

Finally, the social network characteristics of the users active at the centre of the attractors and the databases engaged, as identified by the Fourier analysis, were compared, using an ordinary least squares regression analysis. The aim was to determine whether network positions such as degree centrality, betweenness or particular brokering roles were associated with attractor conditions.

4 Results

The results are presented in five sections, corresponding with the methods discussed above: data trends, Fourier analysis, usage distribution analysis, social network analysis, and regression analysis.

[3] Power-law distributions are common in large networks, underpinning the "small world" phenomena whereby a small number of nodes are highly connected to a large number of small-connectors [33]. This has been theorised as the outcome of "preferential attachment", the tendency for people to link with those who are already well-connected [39]; originally proposed by Price [40]. However, the notable exponential cut-off in many network distributions as in this case has prompted modified explanations employing notions of link-decay or phase-change with age [35-36] or a physical limit to the number of links that can be maintained by a single connection [37]; see also an early formulation by Simon [38]. Fenner, Levine et al. [34] present a modification of the power-law distribution to account for this cut-off: $f(i) = Ci^{-\tau}q^i, 0 < q \leq 1$, i.e. providing a power-law distribution when q = 1 but reducing the exponent, and thus the tail of the distribution, in other cases.

4.1 Usage Trends

Fig. 1 illustrates the adoption of the knowledge base. Rogers' [18] theory of the adoption of innovations posits an 's-curve' pattern as an innovation successively spreads to early adopters, suggesting perhaps that some sections of a population are psychologically disposed to greater experimentation and risk taking than others. While the cumulative proportion of the workforce making some use of the knowledge base is arguably an "s-curve", the pattern can perhaps be more usefully interpreted as an upward phase shift in 2004. This is more apparent in the trend in total usage, which also suggests a second, downward shift in 2006.

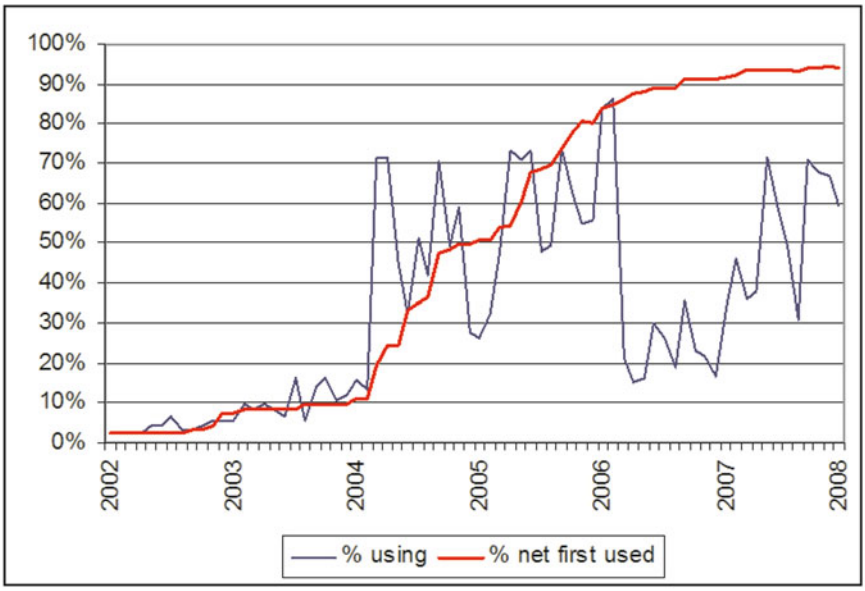


Fig. 1: Adoption of knowledge base by staff (monthly usage)

Table 2 presents data on usage per user. After an initial establishment period, gross usage increases rapidly in 2004 and 2005, falls back in 2006 then nearly doubles in 2007. The mean usage per user increased throughout the period, tripling annually in the first three years then decelerating, reinforcing the notion of the 'phase shift' above, but pointing to intensive and not just extensive use.

Tables 3 and 4 demonstrates that the major use of the depository regarded research and development issues, suggesting some degree of collaboration. But this is eclipsed by postings and viewings of product specifications and general corporate information. Further, postings are largely made by the IT division,

which has responsibility for managing and developing the knowledge base, the corporate division and Division B from 2004. Much of the reading was by the posting divisions, suggesting a usage for internal purposes, while the remaining divisions read more than is posted. Together these tables suggest that the knowledge base is driven by central corporate priorities more than an organic development from users.

Tables 5 and 6 provide additional support for this proposition, with the largest proportion of staff from this division and IT using the knowledge base from establishment. The IT division embraced the tool most comprehensively, with the corporate division being eclipsed by all but one of the divisions in 2005. However by 2005, 14% of the staff had not used the system.

Initial use of the knowledge base centred on the staff directory and on IT and corporate documentation, reflecting the centralised origin of the initiative. The great increase in usage was concentrated on product specifications, quality assurance and procedures, related to standardisation. The third period associated with increased usage in R & D and customer relations, review and planning, pointing to more integral use of the depository for more involved processes.

Table 2: Usage by user

	2002	2003	2004	2005	2006	2007	Total
All							
Total	3639	3744	19287	25612	19102	36540	107924
Users	13	37	119	141	156	178	232
Mean	280	101	162	182	122	205	465
Max	440	325	482	605	919	1255	1255
Std	30	12	20	14	15	21	18
Read							
Total	2454	2338	10036	19403	15506	28097	77834
Readers	12	32	119	141	155	178	232
Mean	205	73	84	138	100	158	335
Max	315	325	311	605	919	1255	1255
Std	24	10	10	12	15	20	15
Posters							
Total	1185	1406	9251	6209	3536	8443	30030
Posters	6	28	62	93	109	74	165
Mean	198	50	149	67	32	114	182
Max	440	169	482	299	350	463	482
Std	18	7	18	6	5	8	9

Minimum is zero in all cases.

Fig. 2 plots the daily usage of the knowledge base, in terms of the number of views and postings. There is an exponential increase in usage during the period, with some evidence of a quarterly periodicity. However, the extensive noise in the data suggests a more complex process than random events around

Table 3: Annual usage by information category and year - transactions

Category	2002	2003	2004	2005	2006	2007	Total
Corporate	588	339	1279	3143	2235	4122	11706
Customer Relations	87	66	624	624	345	2111	3857
Division A		77	457	314	86	150	1084
Division B			169	781	182	2269	3401
Division C	12	87	236	224	157	112	828
Division D		33	108	853	269	181	1444
Division E			120	162	202	612	1096
Export		4	61	52	104	421	642
Franchise	46	468	405	2325	1094	1205	5543
IT	939	304	276	450	1053	1456	4478
Procedures	89	475	2795	3513	1995	3051	11918
Product Specifications	49	24	5783	4956	2605	3223	16640
QA	444	317	3025	3583	2838	3510	13717
R&D	90	396	1990	1925	3569	6479	14449
Review & Planning	7	152	562	1073	1317	3664	6775
Staff Directory	1288	1002	1397	1634	1051	3974	10346
Total	3639	3744	19287	25612	19102	36540	107924

Table 4: Annual usage by information category and year - proportion of transactions

Category	2002	2003	2004	2005	2006	2007	Total
Corporate	16%	9%	7%	12%	12%	11%	11%
Customer Relations	2%	2%	3%	2%	2%	6%	4%
Division A		2%	2%	1%			1%
Division B			1%	3%	1%	6%	3%
Division C		2%	1%	1%	1%		1%
Division D		1%	1%	3%	1%		1%
Division E			1%	1%	1%	2%	1%
Export					1%	1%	1%
Franchise	1%	13%	2%	9%	6%	3%	5%
IT	26%	8%	1%	2%	6%	4%	4%
Procedures	2%	13%	14%	14%	10%	8%	11%
Product & Specifications	1%	1%	30%	19%	14%	9%	15%
QA	12%	8%	16%	14%	15%	10%	13%
R&D	2%	11%	10%	8%	19%	18%	13%
Review & Planning		4%	3%	4%	7%	10%	6%
Staff Directory	35%	27%	7%	6%	6%	11%	10%
Total	100%	100%	100%	100%	100%	100%	100%

a stochastic trend. A six-order polynomial trend line produces an R^2 of only 0.007, for example.[4]

[4] Excel's optimisation algorithm produces a best fit with
y = 2E-19x6 - 5E-14x5 + 5E-09x4 - 0.0002x3 + 6.6262x2 - 101564x + 6E+08

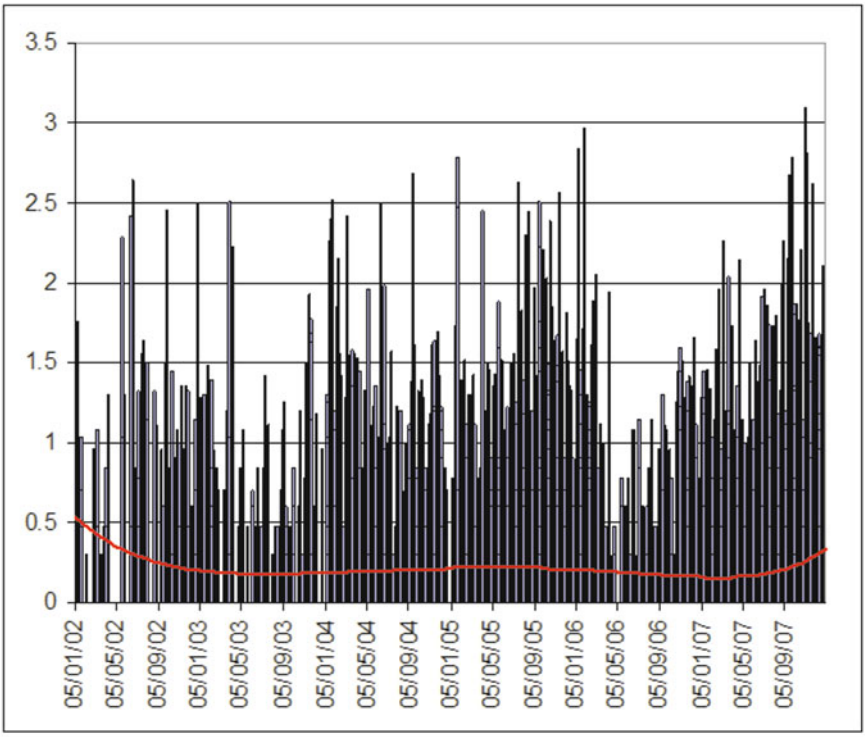


Fig. 2: Daily usage (log)

4.2 Fourier Analysis

By contrast, Fig. 3, plotting change in usage against change in users, suggests a persistent, complex relationship at play, with small changes intermittently having large effects and increasing volatility in activity.

Volatile and discontinuous functions such as those presented in Fig. 3 can be approximated as a Fourier series, the results of which are presented in Fig. 4. This indicates a concentration in the centre of oscillations over time. In 2002 activity centres on a few large oscillations in the sine range 370-410, most likely the initial populators of the databases. In subsequent years the most frequent sine range declines from 110-120 to 60-70 to 50-60.

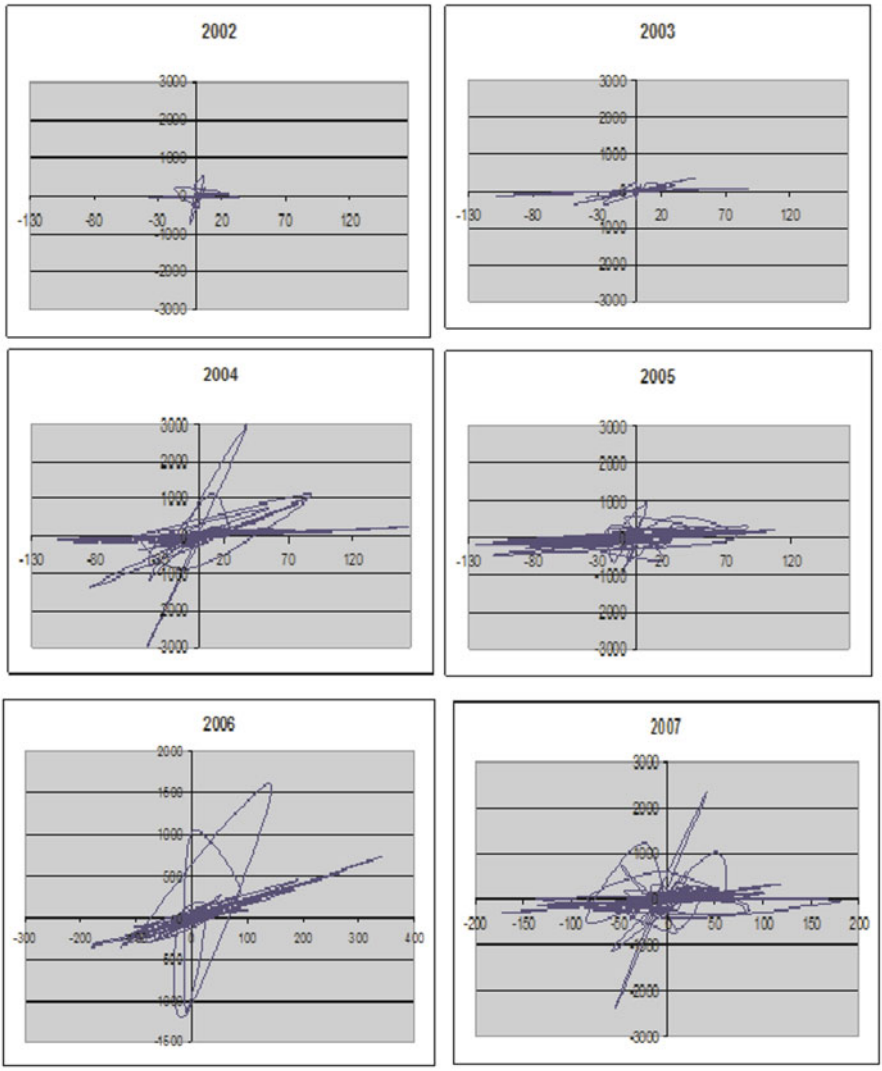


Fig. 3: Annual daily aggregated change in usage(y) with change in users(x)

Table 5: Annual usage by divisional origin.

	2002	2003	2004	2005	2006	2007	Total N	Total %
Total								
Corporate	1570	1421	5016	6049	4666	15542	34264	3%
A		7	816	1615	1794	1284	5516	5%
B		4	1569	6276	3006	7414	18269	17%
C	52	669	793	3229	2682	2065	9490	9%
D		33	865	2241	2322	1909	7370	7%
E			84	85	788	2778	3735	3%
IT	2017	819	10048	6009	3141	5437	27471	25%
Partner		791	96	108	694	40	1729	2%
Not identified					9	71	80	0%
Total	3639	3744	19287	25612	19102	36540	107924	100%
Reading								
Corporate	1293	634	3538	4431	3647	10783	24326	31%
A		6	736	1395	1550	1129	4816	6%
B		2	966	4511	2588	6363	14430	19%
C	52	410	668	2889	2207	1724	7950	10%
D		28	837	2106	2034	1774	6779	9%
E			82	83	631	2073	2869	4%
IT	1109	555	3114	3880	2245	4153	15056	19%
Partner		703	95	108	595	37	1538	2%
Not identified					9	61	70	0%
Total	2454	2338	10036	19403	15506	28097	77834	100%
Posting								
Corporate	277	787	1478	1618	1019	4759	9938	33%
A		1	80	220	240	155	696	2%
B		2	603	1765	413	1051	3834	13%
C		259	125	340	470	341	1535	5%
D		5	28	135	288	135	591	2%
E			2	2	157	705	866	3%
IT	908	264	6934	2129	883	1284	12402	41%
Partner		88	1	0	66	3	158	1%
Not identified						10	10	0%
Total	1185	1406	9251	6209	3536	8443	30030	100%

4.3 Usage Distribution Analysis

Figs. 5 and 6 present the distribution of views and postings per user respectively. This pattern clearly exhibits a power-law distribution with an exponential cut-off.

Table 6: Proportion of divisional staff using knowledge base by year

Division	2002	2003	2004	2005	2006	2007
Corporate	15%	24%	71%	62%	68%	62%
A	0%	8%	57%	70%	76%	81%
B	0%	2%	27%	37%	54%	85%
C	9%	21%	51%	74%	70%	66%
D	0%	11%	52%	78%	76%	78%
E	0%	0%	23%	23%	46%	100%
IT	60%	60%	80%	100%	100%	100%
Partner	0%	58%	67%	25%	25%	25%

4.4 Social Network Analysis

Fig. 7 presents a visualisation of the pattern of aggregate usage, by division, at the start of the study. Views are represented by arrows from each section of the knowledge base and postings by arrows towards each section. This shows the initial flow of data from three staff members, particularly one in IT, establishing the knowledge base primarily around R&D. Table 6 shows the density of usage decreases over time as more users make use of the system and it ceases to be a tool of a small number of intensive users, followed by more widespread take-up form 2006.

Table 7: Evolution of network density

	Average value per possible tie	Standard deviation
2002	19.2	103.2
2003	3.6	44.5
2004	1.0	22.5
2005	1.5	27.5
2006	7.7	45.8
2007	12.1	123.6

Tables 7 and 8 present the evolution of the social network from establishment to 2007. After Everett and Borgatti [41], centrality measures for each mode (databases and users) are calculated separately. It can be seen from Table 7 that the most central databases are increasingly the focus for the network, through. The particular focus shifts, however, from initial research and development to product specification, procedures and corporate concerns, reflecting more widespread consultation of corporate standards. The exception to this trend is the declining eigenvector centrality, reflecting increased consultation of secondary databases alongside the most central one. From 2006 this is also reflected in the reduced centrality of the most central database on other metrics.

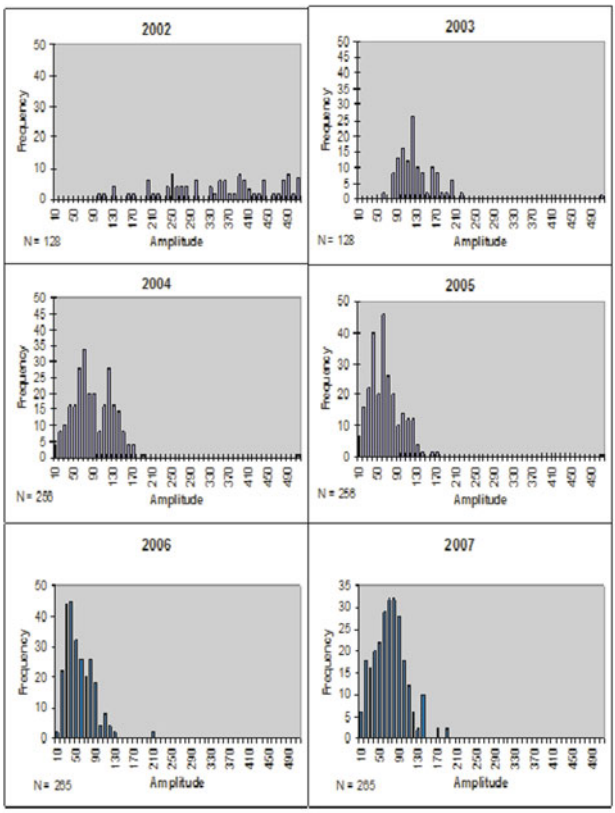


Fig. 4: Fourier spectra - analysis of oscillations in usage per user.

Table 8: Database centrality

	2002	2003	2004	2005	2006	2007
Degree	47.4	36.7	75.8	80.6	47.4	52.6
Closeness	52.8	46.7	70.4	76.4	50.9	25.9
Betweenness	58.9	38.3	35.2	24.1	21.8	30.9
Eigenvector	76.9	58.2	56.7	43.6	44.3	49.3
Central Node	R&D	Product Specs	Procedures	Corporate	Corporate	QA

Table 8 demonstrates the central importance of the main populators of the knowledge base, principally user 37, a major figure in the IT Division. The centrality of the main user declines over time as more users engage with the system. The betweenness and eigenvector centrality of the most

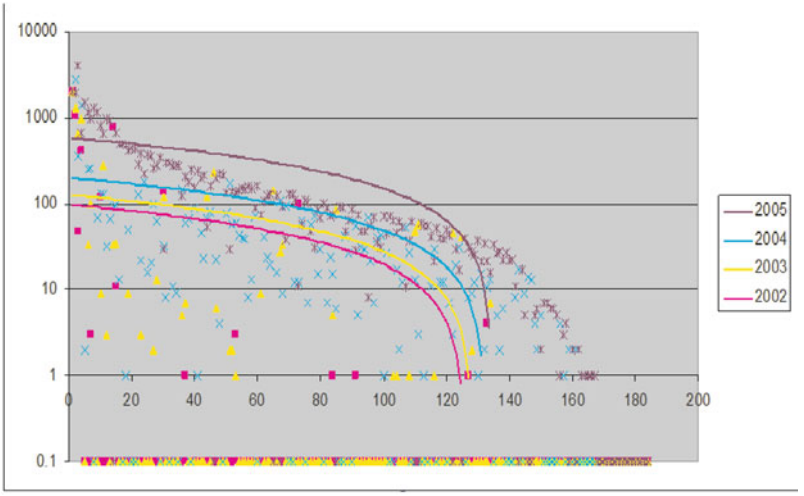


Fig. 5: Annual views per user

central actors increases from 2006, however, suggesting a strategic role of these figures. Finally, as demonstrated in Table 9, an aggregation of the use of the depository suggests that information was increasingly shared through the knowledge base between divisions. This was particularly concentrated on the Corporate and IT Divisions and Divisions C and D.

Table 9: User centrality
a. Betweenness: User 26

	2002	2003	2004	2005	2006	2007
Degree	21.1	14.3	7.0	8.5	7.6	7.3
Closeness	55.9	53.8	51.8	52.2	48.9	25.5
Betweenness	51.7	17.4	1.8	0.74	2.3	2.6
Eigenvector	53.4	41.9	15.0	12.7	17.5	18.5
Central Node	37	37; 28	37; 78	37	139[a]	37

Table 10: Inter-divisional information sharing via the electronic depository

	2002	2003	2004	2005	2006	2007
Density Highest	1.7	14.7	151	194	383	492
Outdegree Highest	IT	Division C	Division C	Division D	IT	IT
Indegree	Corporate	Corporate	Division C	Division C	Corporate	Corporate
	Division C		Division D	Division D		

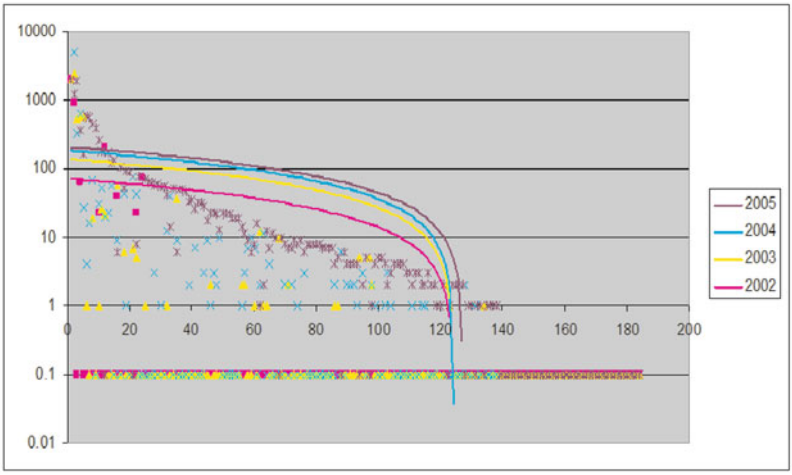


Fig. 6: Annual postings per user

4.5 Regression Analysis

The last stage of the analysis was to identify the social network characteristics of the active nodes at the centre of the attractors, that is, users sharing the most modal amplitude in their pattern of usage each year. An ordinary least squares regression was undertaken of the annual postings and reads of these users within the modal amplitude each year against the four standard indicators of centrality within the information exchange network as a whole for the same year. These were degree, closeness, betweenness and the absolute value of eigenvector centrality. Table 10 summarises the regression results. Two sets of models were significant, the combined pooled cross sectional analysis of all years, and the models for 2004. Looking at the data as a whole, network characteristics account for more than a quarter (26%) of the variation in the attractor activity, that is, the attractors have a major impact on the nature of the information sharing network, particularly reading. The high standardised coefficients for degree centrality indicate that the simple acts of reading and posting by this group are the main influences in constituting the attractors. However, the strong negative correlation with eigenvector centrality indicates that reading and posting by nodes from outside the centre of the network had a greater impact than those at the centre.[5] A modest impact is detectable from closeness, suggesting a group of very active users, consistent with the

[5] While eigenvector centrality frequently presents negative numbers, for the purpose of the regression the absolute value of this indicator was used, so the direction of the association can be interpreted normally.

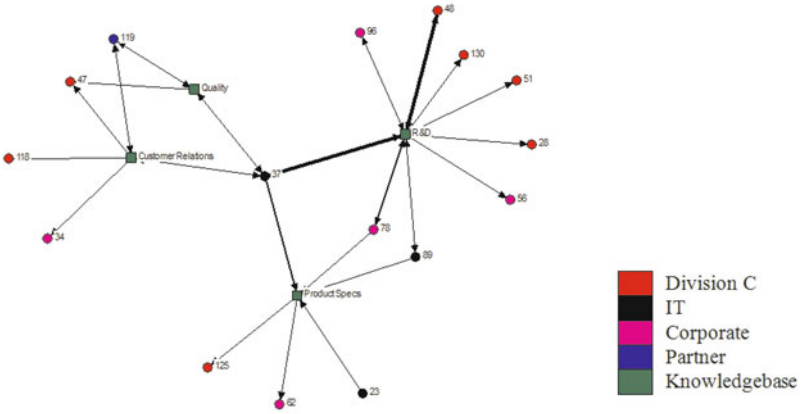


Fig. 7: Social network from pattern of usage 2002

power law finding and the "seeding" role of the IT group. Similarly, the negative correlation with betweenness suggests the information being read and was not at the centre of collaborative interaction across the network.

The results for 2004 show a dramatic exception to these findings. The activity at the centre of the usage that year was highly central to interaction across the network, particularly the postings. Thus, the information posted this year seems to have been responsible for the phase shift in usage of the system.

Table 11: Network characteristics of node activity at the centre of attractors

	All		Read		Post	
2002-07						
Degree	0.851	***	0.836	***	0.663	***
Closeness	0.149	*	0.142	**	0.098	*
Betweenness	-0.123	*	-0.165	*	-0.037	
Eigenvector	-0.503	***	-0.495	***	-0.433	***
Adj R^2	0.261		0.24		0.145	
F (sig)	.000		.000		.000	
2004						
Degree	-0.046		0.300		-0.003	
Closeness	-0.253		-0.245		-0.087	
Betweenness	0.734	***	0.413	**	0.605	**
Eigenvector	0.249		0.146		-0.053	
Adj R^2	0.499		0.39		0.275	
F (sig)	.000		.000		.000	

Standardised Coefficients
Sig: * = .01, ** = .05, *** = .000

5 Discussion and Conclusion

The pattern of activity analysed in the logs of these databases are categorisable as the adoption of an innovation, in Rogers' [18] terms, and the engagement stage of collaborative practice, in Gongla and Rizzuto's [19] scheme. However, such a categorisation reveals only part of the story and there is considerable evidence of social complexity in the manner defined by this study.

Populators appear key to the adoption of this process in general postings at the centre of the attractors and as absent among those who are viewing. A critical shift is apparent in 2004. This was centred on posting and viewing by social brokers, who led more widespread viewing. Thus, from a social complexity perspective, the critical elements or attractors are what is posted and the extent to which brokers are engaged in this in terms of posting and viewing. This interrelationship, framed by the technology, forms the "attractor" at the heart of the complex dynamic process.

The practical implications of this are that in implementing information sharing processes it is critical to identify the posting and viewing needs of brokers and plan their deployment. A careful deployment of information with view to progressively engaging key parts of the organisation is likely to accelerate adoption and consolidate use, providing organisational efficiencies and more effective return on investment than is usually the case in such projects.

This study suggests that social complexity is a useful metaphor for the dynamics of information sharing. The associated concepts of clustered connectivity, great variability and critical attractors are important to making sense of dynamic social networks. The focus on attractors identifies critical conditions for innovation adoption. This contrasts with more passive theories of psychological predisposition or linear accumulation. The study also suggests there is scope for further research on attractor conditions, which are likely to be related to informal networks and discrimination among different types of broking.

Finally, the data collection method is particularly important, providing the potential for continuous analysis of social network evolution in contrast to the norm of comparative cross-sectional analysis. Analysing the readily-collected data on reads and postings provides a valuable gateway into the complex social interactions in information exchange. The techniques employed to identify attractors and their constituents, via Fourier analysis, social network analysis and regression of the two, provides an efficient method for analysing large amounts of data. Such data are typically available in information systems of this sort. Fourier analysis, in particular, is very scalable for use with large datasets at various levels of granularity, whether annually or much more frequently. The social network approach used here, comparative cross-sectional, is less scalable and only viable for annual comparisons. Algorithms for the social network analysis of long series are now available to assist this project.

References

1. Nonaka, I., Takeuchi, H.: The knowledge-creating company: how Japanese companies create the dynamics of innovation. Oxford University Press, New York (1995)
2. Boisot, M.: Knowledge assets: securing competitive advantage in the information economy. Oxford University Press, Oxford (1998)
3. Snowden, D.: Complex acts of knowing: paradox and descriptive self-awareness. Journal of Knowledge Management 6 (2002) 100-111
4. Eagle, N., Pentland, A.: Reality mining: sensing complex social systems. Personal and Ubiquitous Computing 10 (2006) 255-268
5. Csermely, P.: Weak links: stabilizers of complex systems from proteins to social networks Springer, Luxembourg (2006)
6. Monge, P.R., Contractor, N.S.: Theories of communication networks. Oxford University Press, New York (2003)
7. Lin, N., Cook, K., Burt, R.S. (eds.): Social capital: theory and research. Aldine de Gruyter, New York (2001)
8. Williamson, O.E.: The economics of organization: the transaction cost approach. American Journal of Sociology 87 (1977) 548-577
9. Fulk, J., Flanagin, A.J., Kalman, M.E., Monge, P., Ryan, T.: Connective and communal public goods in interactive communication systems. Communication Theory 6 (1996) 60-87
10. Cook, K.: Exchange and power in networks of interorganizational relations. In: Benson, K. (ed.): Organizational analysis. Sage, London (1977)
11. Berger, C.R., Calabrese, R.J.: Some explorations in initial interaction and beyond: toward a developmental theory of interpersonal communication. Human Communication Research 1 (1975)
12. McPherson, J.M., Smith-Lovin, L.: Homophily in voluntary organizations: Status distance and the composition of face to face groups. American Sociological Review 53 (1987) 370-379.
13. DiMaggio, P.J., Powell, W.W.: The iron cage revisited: institutional isomorphism and collective rationality in organisational fields. American Sociological Review 48 (1983) 147-160
14. Burt, R.S.: Structural Holes: The social structure of competition. Harvard University Press, Cambridge, MA (1992)
15. Cross, R., Parker, A., Prusak, L., Borgatti, S.P.: Knowing what we know: Supporting knowledge creation and sharing in social networks. Organizational Dynamics 30 (2001) 100-120
16. Palazzolo, E.T., Serb, D.A., She, Y., Su, C., Contractor, N.S.: Coevolution of communication and knowledge networks in transactive memory systems: using computational models for theoretical development. Communication Theory 16 (2006) 223-250
17. Su, C.: Where to get information in the workplace? A multi-theoretical and multidimensional network analysis on information retrieval from team members and digital knowledge repositories. Sunbelt XXVIII International Network for Social Network Analysis, Tradewinds Resort, St Pete's Beach, Florida (2008)
18. Rogers, E.M.: Diffusion of innovations. Free Press, New York, NY: (1962)
19. Gongla, P., Rizzuto, C.R.: Evolving communities of practice: IBM Global Services experience. IBM Systems Journal 40 (2001) 842-862
20. Afuah, A.: Innovation management. Oxford University Press, Oxford (2003)
21. Roberts, E.B., Fusfeld, A.R.: Staffing the innovative technology-based organization. Sloan Management Review (1981) 19-34
22. Iansti, M.: Real-world R&D. Jumping the product generation gap. Harvard Business Review (1993) 139-147
23. Allen, T.: Managing the flow of technology. MIT Press, Cambridge, MA (1984)

24. Tushman, M., Nadler, D.: Organizing for innovation. Cambridge Management Review 28 (1986) 74-92
25. Schon, D.A.: Champions for radical new inventions. Harvard Business Review 41 (1963) 77-86
26. Howell, J.M., Higgins, C.A.: Champions of technological innovation. Administrative Science Quarterly 35 (1990) 317-341
27. Maidique, M.A.: Entrepreneurs, champions, and technological innovation. Sloan Management Review (1980) 59-76
28. Isherwood, D.: Intel Corporation (UK) Ltd: 10 critical success factors for Notes adoption. In: Little, S., Quintas, P., Ray, T. (eds.): Managing knowledge: an essential reader. The Open University in association with Sage Publications, London (2002) 272-279
29. Ayers, L.R., Beyea, S., Godfrey, M.M., Harper, D.C., Nelson, E.C., Batalden, P.B.: Quality improvement learning collaboratives. Quality Management in Health Care 14 (2005) 234-247
30. McKelvey, B.: Simple rules for improving corporate IQ: basic lessons from complexity science. In: Abdriani, P., Passiante, G. (eds.): Complexity theory and the management of networks: Proceedings of the Workshop on Organisational Networks as Distributed Systems of Knowledge, University of Lecce, Italy, 2001. World Scientific, Imperial College Press, Singapore (2004) 39-52
31. Sawyer, R.K.: Social emergence: societies as complex systems. Cambridge University Press, Cambridge (2005)
32. Gerald, C.F., Wheatley, P.O.: Applied numerical analysis. Pearson, Boston (2004)
33. Watts, D.: Networks, dynamics and the small-world phenomenon. American Journal of Sociology 105 (1999) 493-527
34. Fenner, T., Levene, M., Loizou, G.: A model for collaboration networks giving rise to a power-law distribution with an exponential cutoff. Social Networks 29 (2007) 70-80
35. Amaral, L., Scala, A., Barthélémy, M., Stanley, H.: Classes of small-world networks. Proceedings of the National Academy of Sciences of the United States of America 97 (2000) 11149-11152
36. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of networks with aging of sites. Physical Review E 62 (2000) 1842-1845
37. Mossa, S., Barthélémy, M., Stanley, H., Amaral, L.: Truncation power law behavior in "scale-free" network models due to information filtering. Physical Review Letters 88 (2002) 138701-138701-138701-138704
38. Simon, H.: On a class of skew distribution functions. Biometrika 42 (1955) 425-440
39. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Science 286 (1999) 509-512
40. Price, D.: A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society of Information Science 27 (1976) 292-306
41. Everett, M.G., Borgatti, S.P.: Extending centrality. In: Carrington, P., Scott, J., Wasserman, S. (eds.): Models and methods in social network analysis. Cambridge University Press, Cambridge (2005)

Harnessing Wisdom of the Crowds Dynamics for Time-dependent Reputation and Ranking

Elizabeth M. Daly

Abstract The "wisdom of the crowds" is a concept used to describe the utility of harnessing group behaviour, where user opinion evolves over time and the opinion of the masses collectively demonstrates wisdom. Web 2.0 is a new medium where users are not just consumers, but are also contributors. By contributing content to the system, users become part of the network and relationships between users and content can be derived. Example applications are collaborative bookmarking networks such as del.icio.us and file sharing applications such as YouTube and Flickr. These networks rely on user contributed content, described and classified using tags. The wealth of user generated content can be hard to navigate and search due to difficulties in comparing documents with similar tags and the application of traditional information retrieval scoring techniques are limited. Evaluating the time evolving interests of users may be used to derive quality of content. In this chapter, we explore a technique to rank documents based on reputation. The reputation is a combination of the number of bookmarkers, the reputation of the bookmarking user and the time dynamics of the document. Additionally, this reputation measure is extended to take into account the time-dependent, term-dependent reputation of a document. Experimental results and analysis are presented on a large collaborative IBM bookmarking network called Dogear.

1 Introduction

Web 2.0 is a new medium where users are not just consumers, but are also contributors. By contributing content to applications, such as Flickr or del.ic.ious, users become part of the network and relationships between

Elizabeth M. Daly
IBM, Dublin Software Lab e-mail: Elizabeth_Daly@ie.ibm.com

users and content can be derived. These social networks between people and content can potentially be harnessed to capture "the wisdom of the crowds" [10]. Example applications are collaborative bookmarking networks such as del.icio.us[1] and file sharing applications such as YouTube[2] and Flickr[3]. These networks rely on content being classified by user generated tags and thus provide a mechanism for informal catagorisation. The rating and ranking of documents in such networks is problematic. Search engines determine rank based on indexable content, in a tagging environment indexable content may be scarce. Google's page rank uses the connectivity of the entire network to influence rank. However, in a tagging based environment content may be an isolated piece of data, with little or no connectivity network between data sources. Additionally, the wealth of user generated content can be hard to navigate and search due to difficulties in comparing documents with similar tags.

One solution is to take into account the popularity of a document. A document that has been consumed by a large number of users can be deemed popular where a 'boost count' denotes the number of consuming users. However, this technique inevitably favours older documents. As a result, newly added content cannot compete with documents consumed by a large number of users in the past. Consequently, ranking based on popularity can lead to a 'rich-get-richer' scenario suppressing newly added documents. Cho and Roy demonstrated that ranking based on popularity hinders the discovery of new web pages and that increases in web page popularity are heavily influenced by search engine ranking [3]. To overcome this bias towards older documents, some applications allow ranking based on recency, where content is ranked based on how recently the content was added to the network. However, ranking based on recency neglects the quality of a document and makes discovery of relevant and reputable content difficult.

Here we explore a technique to rank documents based on reputation, illustrated in figure 1. In order to harness the "wisdom of the crowds" the reputation of users and the reputation of documents are integrated into a single metric. Consequently, reputable users, e.g., trend-setters have a greater influence on the reputation of bookmarked documents than users with a low reputation. To address the age bias, both reputation metrics undergo a decay process. This means that inactive users and documents decrease in reputation over time with the advantage of capturing the current trends. Reputation in this case is treated as global, where the strength of a term association with a document is ignored. A concern with this approach is that as more users annotate a document, the more noisy the text associated with the document may become. As a result, we extend this reputation measure to include term specific reputations between users, documents and terms. Experimental re-

[1] www.delicious.com

[2] www.youtube.com

[3] www.flickr.com

sults and analysis are presented on a collaborative IBM bookmarking network called Dogear.

2 Social and Term Ranking Based on Reputation

A Reputation Ranking is proposed which captures the time dynamic reputation of a bookmarked document. This work is inspired by a Delay-Tolerant Network (DTN) routing algorithm which measures the predictability of mobile node encounters [8]. Four factors are integrated into the reputation of a document:

1. The number of users consuming the document;
2. The reputation of the user contributing the document;
3. The time dynamics of user consumption;
4. The time dynamics of consumption of documents contributed by the user.

Incorporating these factors into a ranking system and coupling the popularity of documents along with the timeliness of recent bookmarking activities shifts the search results from mainly dominated by old results to more recent up-to-date results. The advantage of such an approach manifests itself in the discovery of trends, which otherwise is more difficult to achieve. Finally, a term specific reputation is maintained on a per-document/per-term basis and a per-user/per-term basis, which rewards documents where the term is frequently applied, and rewards users who apply terms to documents that are then subsequently reapplied.

2.1 User Reputation

Users consume content contributed to the network by other users. The number of users consuming a given contributors content may be seen as a form of implicit recommendation. If a user consistently adds content that other users

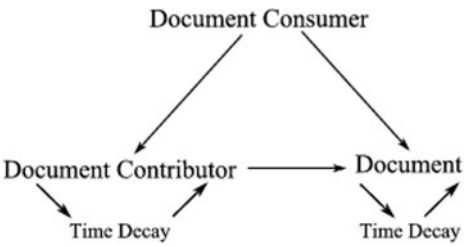


Fig. 1: Reputation Network

deem interesting, then this user can be considered a reputable contributor. The reputation of a user is initialised to an input parameter $R_{init} \in [0, 1]$. After this time, the value is updated with a reward constant R_{reward}

$$R_{new} = R_{old} + (1 - R_{old}) \times R_{reward}. \quad (1)$$

Equation 1 is applied to a user every time a person adds a document to their collection contributed by that user. The selection of the reward constant R_{reward} should consider the consumer rate of the application. If the number of expected consumers is in the order of hundreds or thousands, then an overly high value of R_{reward} will potentially cause popular content to quickly converge towards 1 making it difficult to differentiate between similarly popular content.

2.2 Document Reputation

The number of users that add a document to their collection indicates popularity and may be used to derive the quality of the document. A simple reward model is used to measure a document reputation. Each time a user adds a bookmark to their library, the reputation of the document is updated, reinforcing the value of the bookmark. As such, a bookmark's reputation can be measured in a similar manner to how user's reputation is measured by applying equation 1.

2.3 Time Dynamics

Golder and Huberman researched bookmark popularity and reported that 67% of pages reached their peak popularity levels in the first 10 days after being added to del.icio.us [4]. However, 17% took over 6 months in order to reach peak popularity.

In order to capture the time dynamics and current relevance of the document to the user population, the reputation value is decayed over time. Therefore, documents that are continuously being added to user libraries are rewarded, and inactive content is degraded by:

$$R_{new} = R_{old} \times \gamma^k, \quad (2)$$

where γ is the decay coefficient and k is the number of elapsed time units since reputation value was last aged. As a consequence, bookmarks with a high reputation are decayed over time unless user consumption remains steady. The selection of the decay coefficient, γ, and time unit, k, should be based on the importance of recency in the application. An aggressive

decay model can be used to detect short term trends with a low value of γ and a time unit of days or even hours. An application where recency is less important than the quality of content, a more conservative decay model may be employed with a value of γ tending towards 1 and a time unit of weeks or months. As with a document reputation, a user's reputation must be time dependent in order to reflect the recent nature of their contributions. As such, user reputations are decayed over time using equation 2 also.

2.4 Reputation Ranking: Combining Document and User Reputation

As shown by Golder and Huberman [4], documents can differ greatly in the rate of user consumption. Evaluating the quality of newly contributed content is problematic when relying on a simple boost count. As a consequence, the proposed solution of Reputation Ranking given in equation 3 utilises the reputation of a document in conjunction with the reputation of the user consuming the document. The Reputation Ranking of a bookmark document is denoted as:

$$R_{new} = R_{old} \times R_{bookmarker} \times \beta, \tag{3}$$

where β is a weighting constant representing the extent to which a consuming bookmarker's reputation may influence a bookmark's reputation. The value of β determines the amount of influence a user's reputation may have over a document's reputation value.

2.5 Term-reputation Ranking: Combining Reputation Ranking and Term Ranking

The Reputation Ranking scheme presented above attempts to evaluate the quality of a given document and does not take into account the text terms users apply when tagging or assigning a title to the document. As a result, ranking purely based on reputation may lead to high quality documents, however, the relevance to the given search term may be weak. Therefore we propose that each user and document maintains a general reputation value as describe above, and also a set of Term-Reputations. The Term-Reputation is calculated in a similar manner to the reputation value given in equation 3, however, a separate reputation value is maintained for each term associated with a document, and each term applied by a bookmarker. The Term-Reputation for a given document d and a given term t is rewarded every time a user annotates the document with the given term.

$$R_{termRep}(d,t) = R_{bookmark}(d) + R_{term}(t) \quad (4)$$

Similarly a Term-reputation for a user u for a given term t is rewarded everytime an additional user annotates a document with term t, if the user u was the first to apply this term, i.e. the term contributor to that document. As with the other reputation values, this value decays over time.

3 Experiment Results

In this section the experiment used to evaluate Reputation Ranking is described and the performance is compared to basic boost ranking.

Table 1: Dogear dataset

Number of Users	Number of Bookmarks	Number of URLs
10259	505472	317362

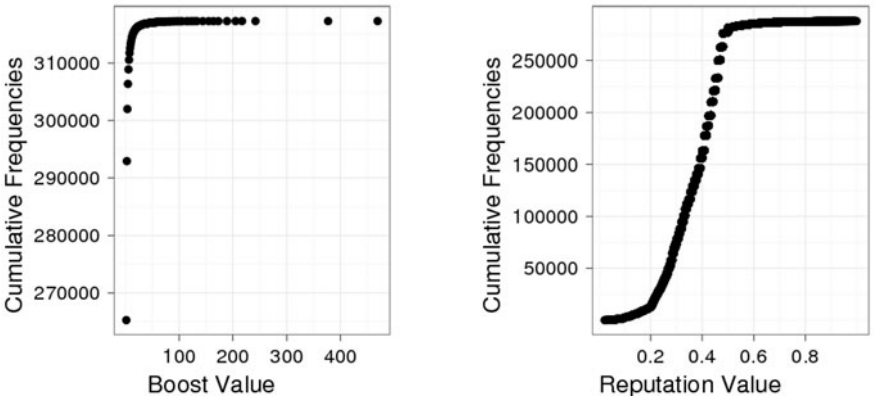
3.1 Experimental Setup

In order to evaluate the premise of Reputation Ranking a large collaborative bookmarking data set is used. IBM's collaborative bookmarking solution Dogear [9] is popularly used by IBM employees. The Dogear data set contains an extensive network of users and contributed URLs, shown in table 1. The time dependent additions of bookmarks to the network are used to calculate the bookmark and bookmarker's reputation. This is achieved by simply replaying the bookmarking and tagging activities in chronological order.

Figure 2 shows the cumulative distribution of ranking based on boost of the entire document collection. As can be seen the graph obeys the power law with a power law coefficient of 2.46 which results in a very long tail with a small number of highly ranked documents.

Upon inspection 99% of the documents have been bookmarked by 13 people or less. As a result, the majority of documents are differentiated by only a small number of additional users bookmarking the document and therefore are difficult to compare. The rank distribution using reputation is cumulative normal which results in a clear divide between highly reputable documents and documents with low reputation.

In order to evaluate the benefits of Reputation Ranking the result set is examined compared to boost ranking. The following details are evaluated:



(a) Boost Ranking (b) Reputation Ranking

Fig. 2: Cumulative distribution of ranked collection

- Boost count: This is a good representation of the popularity of a document, however, evaluating based on that alone rewards a rich-get-richer ranking algorithm and needs to be viewed in the context of document recency also.
- Time since last modified: A document with high boost count, though popular, may not be highly relevant if a long period of time has passed since the document has been active in the network, i.e., the document has been bookmarked.
- Document life-span: This is the difference between the time the document has last been active compared to when a document was first added to the network. Boost tends to favour documents that have been in the network for long periods of time, Reputation Ranking on the other hand aims to reduce this life-span dependency.
- Distribution of person reputation: In order to reduce the life-span dependency, other measures need to be taken into account in order to determine the quality of a newly contributed document. This is achieved through the additional reward of documents bookmarked by users with a high reputation.

In order to demonstrate the utility of Term-Reputation Ranking we evaluate the extent to which a term has been applied to describe the high ranking documents using each ranking mechanism.

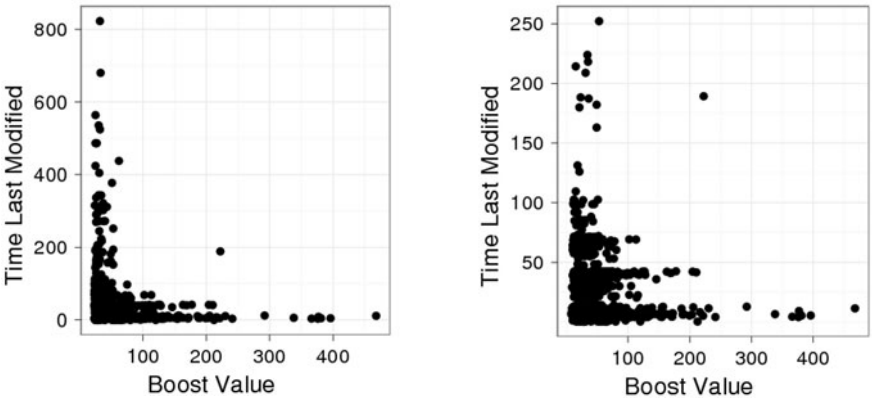
Table 2: Reputation Rank parameter values

R_{init}	R_{reward}	γ	k	β
0.5	0.1	0.98	number of months	0.2

Table 2 shows the parameter values used to calculate the reputation of documents and users. Each document and user is initialised to a value of 0.5, i.e., no user or document is considered to have a different ranking until either user uptake or time determine otherwise. A relatively low decay model was chosen where the time unit of decay is months.

3.2 Document Ranking

In order to demonstrate the time dependent benefits of reputation ranking figure 3 shows the time since the document was last added to the network, compared to ranking based on boost value of the top 1000 documents in the network. Boost ranking returns documents that have not been updated in 2 years. Reputation ranking returns significantly more recent results, while still returning documents with a high boost count that have been recently updated. The overlap of results returned is 63%. Older results with a high boost value, have been replaced by results with a lower boost value but that have been added recently.

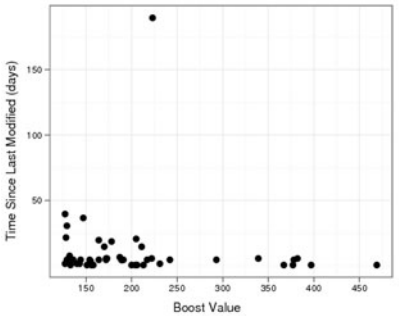


(a) Boost Ranking (b) Reputation Ranking

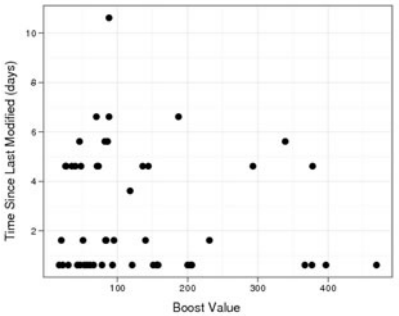
Fig. 3: Ranking of the top 1000 documents

3.3 Ranking Popular Documents

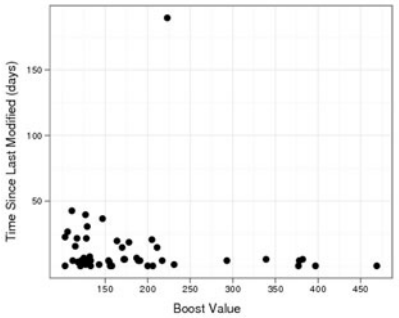
In order to demonstrate ranking highly popular documents, the most frequently used tags across the document collection were selected that return the most results. Figure 4 shows the top 50 results returned based on a tag search for the two most popular tags, in this case "ibm" and "web2.0". The trend clearly shows that ranking based on boost returns a number of older bookmarks when compared to Reputation Ranking. Upon analysis, the search results returned by the ranking schemes have an overlap of 38% and 44% for "ibm" and "web2.0" respectively. The overlap is low, because the maximum time since a document was last modified is 11 days for Reputation Ranking and as high as 190 days in boost ranking. This highlights the fact that Reputation Ranking includes more than 50% recent documents that were ignored by boost count ranking.



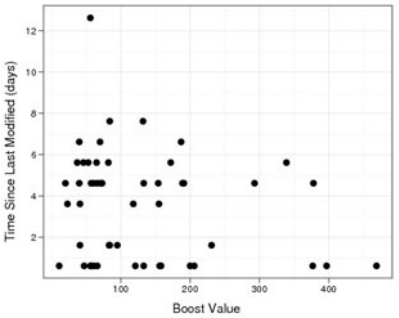
(a) Boost Rank for search term "ibm"



(b) Reputation Rank for search term "ibm"

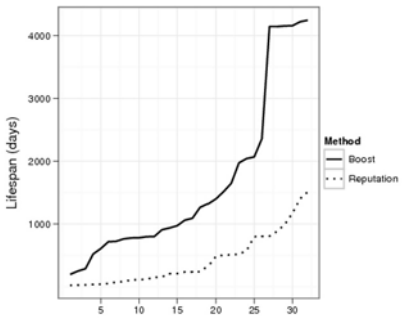


(c) Boost rank for search term "web2.0"

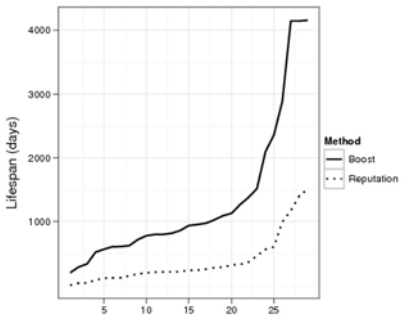


(d) Reputation rank for search term "web2.0"

Fig. 4: Top 50 ranked document



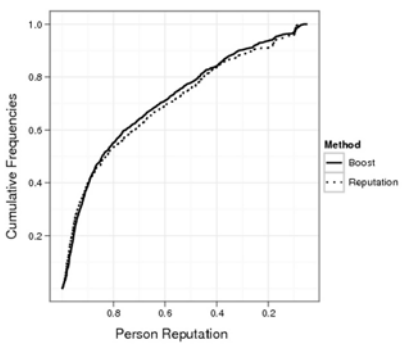
(a) search term "ibm"



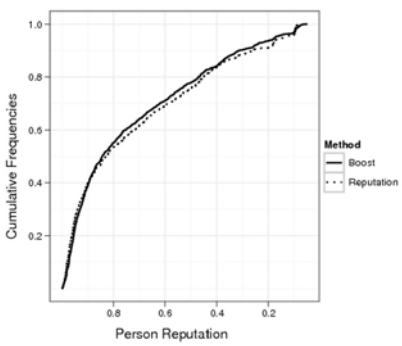
(b) search term "web2.0"

Fig. 5: Age of documents not overlapping in results

As previously discussed, ranking based on boost count significantly favours older documents with a long life-span. In order to evaluate this observation, the results of both ranking mechanisms are examined and only the non-overlapping documents are analysed. Figure 5 shows the life-span of the non-overlapping results of both Reputation Ranking and boost ranking. In both cases, boost rank has returned documents with a high life-span, whereas Reputation Ranking returns documents with a shorter life span but that have been updated more recently.



(a) search term "ibm"



(b) search term "web2.0"

Fig. 6: Person Reputation distribution in non overlapping results

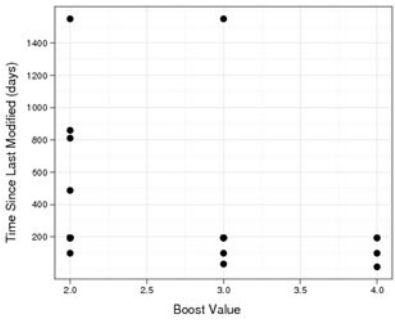
3.4 Ranking Less Popular Documents

As was seen in figure 2, 99% of the documents in the network are bookmarked by 13 people or less. As a result ranking based on a simple boost count has limited meaning as there is no clear divide between the document ranks. If 200 documents have been bookmarked 3 times, the selection of the top 50 results is arbitrary. In order to demonstrate the ranking behaviour, the result set is limited to documents that have been bookmarked 13 times or less. This is achieved by selecting the two most frequently used tags that occur exclusively in documents that have been bookmarked 13 times or less, where more than one person has applied the tag [4]. In this case the two terms are "backpacking" and "photoshop". The disadvantage of ranking based on boost count is most obvious when there is little to differentiate results once the boost count drops to 2. Additionally, some results are as old as 4 years which are clearly not current and may no longer be relevant. Reputation Ranking on the other hand strikes a balance between recent additions and frequently consumed documents.

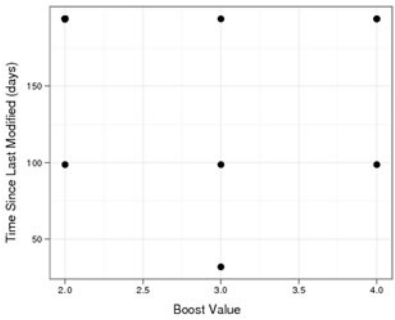
Overlap for the search term "backpacking" is 82% and so there are only a few outliers that are included in the non-overlapping result set shown in figure 8. The age of the documents returned by Reputation Ranking are all zero, as such, the reputation of the user consuming the document has been a deciding factor in ranking these results. The overlap for the search term "photoshop" is 46%.

Figure 9 shows the distribution of the reputation of users that bookmarked the documents in the non-overlapping results set. In figure 9 a) the reputation of the contributing users in Reputation Rank are high, which supports the observation that the reputation of documents with a short life-span are influenced by the users who contributed the documents. The reputation of contributing users included in the boost count distribution are also high while including a small number of users with low reputation. In figure 9 b) the maximum reputation of contributing users is slightly higher in boost rank compared to Reputation Rank. However, the overall reputation of the contributing users are higher for Reputation Rank meaning that a number of users with an above average reputation combined can have a higher influence than a small number of highly reputable users.

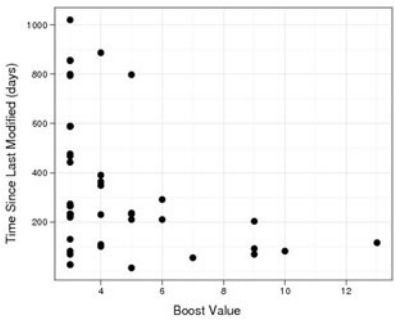
[4] We discarded two tags that were used by a single author to tag all content in their collection. The results showed identical behaviour, however we felt the utility of searching based on a personal tag was limited to that user.



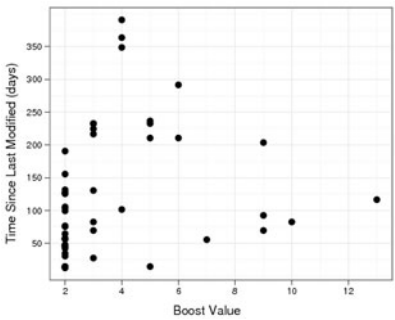
(a) Boost Rank search term "backpacking"



(b) Reputation Rank search term "backpacking"

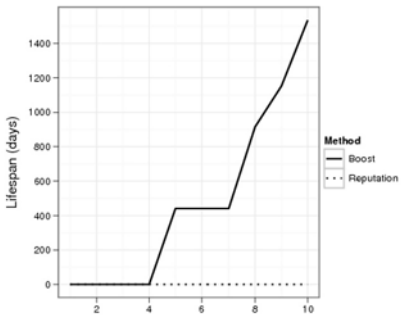


(c) Boost Rank search term "photoshop"

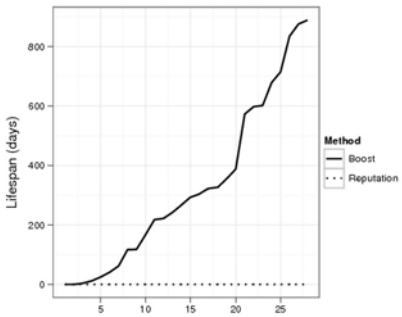


(d) Reputation Rank search term 'photoshop'

Fig. 7: Tag search for most frequently used tag where document bookmarked 13 times or less

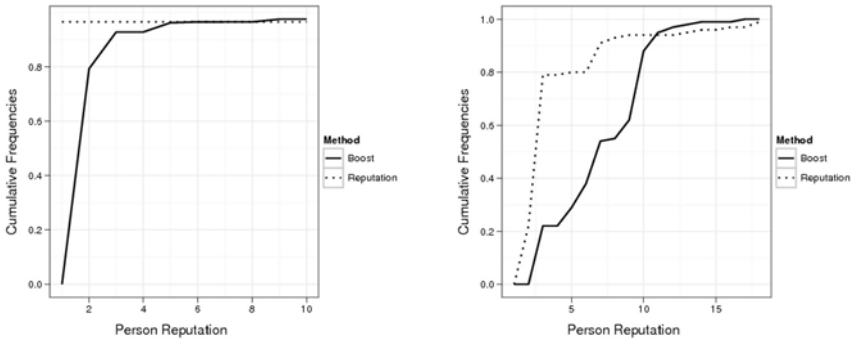


(a) search term "backpacking"



(b) search term "photoshop"

Fig. 8: Age of non-overlapping results



(a) search term "backpacking" (b) search term "photoshop"

Fig. 9: Person Reputation Distribution of non-overlapping results

3.5 Ranking Popular Documents Using Term-reputation

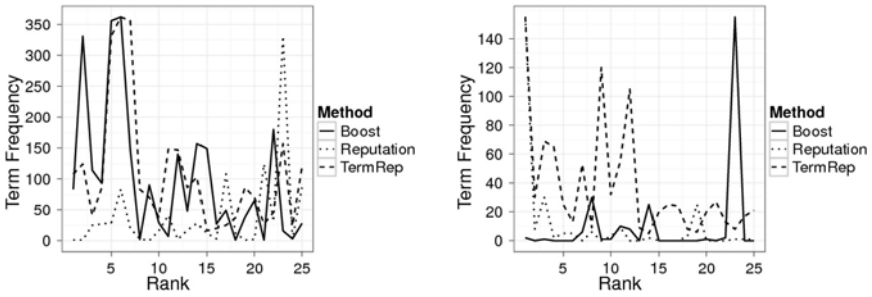
The previous sections have demonstrated the value of Reputation Ranking. The results have been evaluated based on recency and popularity along with the reputation of the users consuming the content. As discussed in section 2.5, Boost Ranking and Reputation Ranking do not take into account the extent to which the results match a given search term, the presence of the search term has been treated as binary. When comparing popular documents that have been annotated by a large number of users, many terms may have been applied to describe the content. As a result, a large number of documents may be returned where the term has little relationship to the content.

Figure 10 shows the term frequency of the given search term compared to document rank position for each ranking mechanism. Figure 10 a) shows that relatively low term frequencies documents are ranked highly by Reputation Ranking compared to Boost Ranking. Boost Ranking returns a number of documents with high term frequencies in the top results, in contrast to Reputation Ranking where the document with the highest term frequency is ranked 23rd. Term Reputation, however, performs better by boosting the documents with a high term frequency improving the overall relevance of the results. Figure 10 b) shows "web2.0" ranked results and surprisingly both Boost Rank and Reputation Rank perform very poorly, where the top results contain very few instances of the search term. Term Reputation ranking performs much better, where the top ranked results have a much higher term frequency.

The poor performance of both Reputation Ranking and Boost Ranking can be explained when examining table 3. This table shows the number of overlapping documents in the top 50 search results returned by each ranking mechanism when searching for the two terms "ibm" and "web2.0". As can be seen Boost Rank performs very poorly where 74% of the results returned when searching for "ibm" are also returned when searching for "web2.0". Reputation Ranking results in a 54% overlap, and though this is an improvement it illustrates the problem that a number of popular results, where a large number of terms are associated with the document may easily drown out more relevant content. Term Reputation shows a significant improvement where there is only a 12% overlap, illustrating the importance of taking into account the relationship between a document and individual terms applied to describe it.

Table 3: Number of overlapping top 50 results

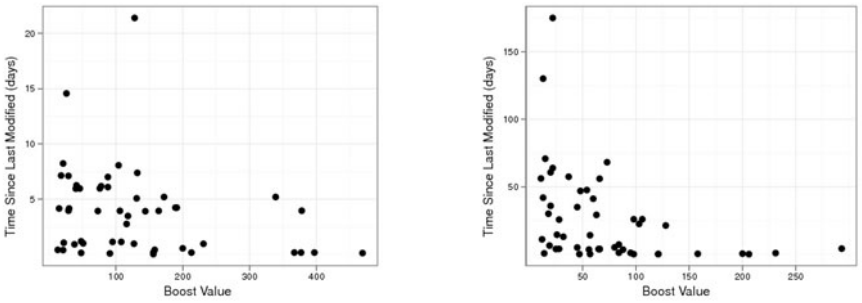
Rank Method	Result Overlap
Boost	37
Reputation	27
TermRep	6



(a) search term "ibm" (b) search term "web2.0"

Fig. 10: Term Frequency vs. Document Rank

Figure 11 shows the time since the document was modified compared to ranking based on Term Reputation of the top 50 results. Figure 11 a) shows that a small number of older documents are included in the search results when compared to those returned by Reputation Ranking, but still provides a trade off of more recent documents when compared to Boost Ranking. Figure 11 b) shows that the results for "web2.0" have excluded a number of documents with a high boost count, due to their weak relationship with the search term. This is in contrast to 4 d) where Reputation Ranking alone returned more documents with a high boost count, but a low term frequency.



(a) TermRep Rank search term "ibm" (b) TermRep Rank search term "web2.0"

Fig. 11: Top 50 ranked document

4 Related Work

Social networks of the Web 2.0 content has been the subject of much recent research on mining social relationships and, most prominently, relations among tags.

Heymann et al. examined social bookmarking and tagging of data in del.icio.us [6]. The authors evaluate the utility of social bookmarking data used to augment Web search. They found that only 20% of the tags do not occur in page text or title of the pages they represent. However, they did show that social bookmarking systems provide a good reflection of changes within the underlying network, illustrating the dynamic nature of content and popularity.

Zaihrayeu et al. attempted to calculate the trustworthiness of search results [12]. Yanbe et al. recently developed a new page re-ranking system using social bookmark information [11]. Bao et al. propose SocialSimilarityRank which

measures the similarity between tags and SocialPageRank which accounts for the popularity among taggers in terms of a frequency ranking [1]. Hotho et al. suggest another variation on PageRank, FolkRank, which is used to improve efficient searching via personalised and topic-specific ranking within the tag space [7]. Zanardi and Capra present a technique to re-rank results utilising the tags most associated with a given user retrieval accuracy for searches based on popular tags [13]. They also take into account the problem of searching documents based on less popular tags by expanding the user query to include tags that are found to be similar based on co-occurrence.

Golder and Huberman provides analysis of the tagging behavior and tag usage in online communities [5]. The authors provide an overview about the structure of collaborative tagging systems. Based on a small subset of the del.icio.us corpus, they investigate what motivates tagging and how tagging habits change over time. Chi and Mytkowicz investigate the dynamics of tags related to documents and shows that the information gained from a tag becomes less useful as the proliferation of use increases [2]. Based on these research results, it can be derived that time dynamics play a key role in the utility of social based ranking schemes.

5 Conclusion

This chapter has presented Reputation Ranking which measures the overall popularity of a bookmark, taking into account the timely relevance of the document. The ranking mechanism presented here is relatively simple, and involves little computational overhead by using a basic reward/decay model. Traditional search methods apply rich-get-richer semantics, which favour old documents in the system and therefore are not able to react to upcoming trends quickly. To address this, the discriminating factor in ranking search results is not the boost count, i.e., the number of times a document has been bookmarked already, but instead a reputation metric. This technique couples the reputation of users with the reputation of the document being bookmarked in an attempt to capture the "wisdom of the crowds". As a result, reputable users, e.g., trend-setters have a greater influence on the reputation of the bookmarked documents than users with a low reputation value. Additionally, reputation is decayed over time only to be reinforced, if documents and users are constantly active. Therefore the recency of a document is an important factor when ranking for given search terms. Documents that once were popular and collected a large amount of bookmarks but have not been used recently will degrade over time. Differentiation between low-ranked documents is problematic when using traditional methods such as boost ranking. In contrast, Reputation Ranking provides a much more fine-grained ranking metric integrating a reward/decay model for both documents and users.

This novel mechanism has been extended to individual documents and users on a term by term basis, where the reputation is not just global, but term specific. In this manner, we have presented Term Reputation Ranking in an attempt to capture a combined measure of content quality, and term relevance. This becomes particularly useful in the case described by Chi and Mytkowicz where tag proliferation, eventually ends up degrading the meaning of the tag [2]. Heymann et al. noted that tags were rarely applied to documents that did not contain the tag text, as a result, the mere presence of a tag is not necessarily a good representation of the extend to which the tag captures the underlying content [6]. By applying time based reputation, older tags, that may no longer be relevant, can be decayed and the presence of a tag for a corresponding document is no longer binary, but a quality measure of the tag can be harnessed.

References

1. Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, and Zhong Su. Optimizing web search using social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 501–510, New York, NY, USA, 2007. ACM Press.
2. Ed H. Chi and Todd Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 81–88, New York, NY, USA, 2008. ACM.
3. Junghoo Cho and Sourashis Roy. Impact of search engines on page popularity. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 20–29, New York, NY, USA, 2004. ACM Press.
4. Scott A. Golder and Bernardo A. Huberman. The structure of collaborative tagging systems. *Technical report, Information Dynamics Lab, HP Labs*, Aug 2005.
5. Scott A. Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208, April 2006.
6. Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can social bookmarking improve web search? In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 195–206, New York, NY, USA, 2008. ACM.
7. A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *In Proceedings ESWC, 2006*.
8. Anders Lindgren, Avri Doria, and Olov Schelén. Probabilistic routing in intermittently connected networks. In *Proceedings of the First International Workshop, SAPIR 2004, Fortaleza, Brazil, August 1-6, 2004*, volume 3126 of *Lecture Notes in Computer Science*, pages 239–254. Springer-Verlag GmbH, August 2004.
9. David R. Millen, Jonathan Feinberg, and Bernard Kerr. Dogear: Social bookmarking in the enterprise. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 111–120, New York, NY, USA, 2006. ACM Press.
10. James Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* New York: Doubleday, 2004.
11. Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, and Katsumi Tanaka. Can social bookmarking enhance search in the web? In *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, pages 107–116, New York, NY, USA, 2007. ACM Press.

12. Ilya Zaihrayeu, Paulo P. da Silva, and Deborah L. Mcguinness. Iwtrust: Improving user trust in answers from the web. In *Lecture Notes in Computer Science*, pages 384–392, 2005.
13. Valentina Zanardi and Licia Capra. Social ranking: uncovering relevant content using tag-based recommender systems. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, pages 51–58, New York, NY, USA, 2008. ACM.

Part III
Discovering Structures in Social Networks

Detecting Communities in Social Networks Using Local Information

Jiyang Chen, Osmar R. Zaïane, and Randy Goebel

Abstract Much structured data of scientific interest can be represented as networks, where sets of nodes or vertices are joined together in pairs by links or edges. Although these networks may belong to different research areas, there is one property that many of them do have in common: the network community structure. There has been much recent research on identifying communities in networks. However, most existing approaches require complete network information, which is impractical for some networks, e.g. the World Wide Web or the cell phone telecommunication network. Local community detection algorithms have been proposed to solve the problem but their results usually contain many outliers. In this paper, we propose a new measure of local community structure, coupled with a two-phase algorithm that extracts all possible candidates first, and then optimizes the community hierarchy. We also propose a community discovery process for large networks that iteratively finds communities based on our measure. We compare our results with previous methods on real world networks such as the co-purchase network from Amazon. Experimental results verify the feasibility and effectiveness of our approach.

1 Introduction

Many datasets can be represented as networks composed of vertices and edges, including the World Wide Web (WWW), organization structures [35],

Jiyang Chen
Department of Computing Science, University of Alberta, e-mail: jiyang@cs.ualberta.ca

Osmar R. Zaïane
Department of Computing Science, University of Alberta, e-mail: zaiane@cs.ualberta.ca

Randy Goebel
Department of Computing Science, University of Alberta, e-mail: goebel@cs.ualberta.ca

academic collaboration records [23, 34] and even political elections [1]. A community in the network can be seen as a subgraph such that the density of edges within the subgraph is greater than the density of edges between inside and outside nodes [15]. The ability to identify communities could be of significant practical importance. For example, groups of web pages that link to more web pages in the community than to pages outside might correspond to sets of web pages on related topics, which could enable search engines to increase the precision and recall of search results by focusing on narrow but topically-related subsets of the web [11]; groups within social networks might correspond to communities, which can be used to understand organization structures. Moreover, the influence of the community structure may reach further than these: a number of recent results suggest that networks can have properties at the community level that are quite different from their properties at the level of the entire network, so that analysis that focus on whole networks and ignore community structure may miss many interesting features [26]. For example, we may find that people in different community groups have different mean numbers of contacts in some social networks, i.e, individuals in one group might have many neighbours while members of another group are more reticent. Such social networks are reported in [2] and [13] for the study of HIV in sexual contact networks. Therefore, characterizing such networks by only quoting a single figure for the average number of contacts an individual has, and without considering the community structure, will definitely miss important features of the network, which is relevant to questions of scientific interest such as epidemiological dynamics [17].

The problem of finding communities in social networks has been studied for decades. Recently, several quality metrics for community structure have been proposed [25, 28, 37]. Among them, modularity Q is proved to be the most accurate [7] and has been pursued by many researchers [6, 10, 16, 26, 36]. However, most of those approaches require knowledge of the entire graph structure to identify communities, which we call *global communities*. A global community is a community defined based on global information about the entire network. That is, one needs to access and see the whole network information. This constraint is problematic for networks which are too large to know completely, e.g., the WWW. In spite of these limitations, finding communities, which we call *local communities*, would still be useful, albeit constrained by the small volume of accessible information about the network in question. A local community is a community defined based on local information without having access to the entire network. For example, we might like to quantify local communities of either a particular webpage given its link structure in the WWW, or a person given his social network in Facebook. Existing approaches [28, 6] also assume that each entity belongs to only one community, however in the real world one entity usually belongs to multiple communities, e.g., one researcher could publish in both the data mining community and the visualization community. We refer to these as overlapping communities.

Several techniques [3, 4, 5, 22] have been proposed to identify local community structure given limited information about network. However, parameters that are hard to obtain are usually required, such as the community size or density. Moreover, communities discovered by these algorithms include many outliers, which are nodes that are weakly connected to the community, and thus suffer from low accuracy. In this paper, we propose a new metric, which we call L, to evaluate the local community structure for networks in which we lack global information. We then define a two-phase algorithm based on L to find the local community of given starting nodes. Moreover, we propose a community discovery process to discover overlapping communities in a large network where global information is not available. Given one or a set of start nodes, our algorithm starts from a local community, then iteratively identifies communities while expanding to the whole graph. We compare our algorithm's performance with previous methods on several real world networks. In contrast to existing approaches, our metric L is robust against outliers. The proposed algorithm not only discovers local communities without an arbitrary threshold, but also determines whether a local community exists or not for certain nodes. Our iterative community discovery process is able to discover overlapping communities with only local information. Additionally it does not require any arbitrary thresholds or other parameters.

The rest of the paper is organized as follows. We discuss related work in Section 2. Section 3 defines the problem and reviews existing solutions. We describe our approach in Section 4 and report experimental results in Section 5, followed by conclusions in Section 6.

2 Related Work

Traditional data mining algorithms, such as association rule mining, supervised classification and clustering analysis, commonly attempt to find patterns in a data set characterized by a collection of independent instances of a single relation. However, for social networks, where entities are related to each other in various ways, naively applying traditional statistical inference procedures, which assume that instances are independent, can lead to inappropriate conclusions about the data [18]. For example, for a search engine, indexing and clustering web pages based on the text content without considering their linking structure would definitely lead to bad results for queries. The relations between objects should be taken into consideration and can be important for understanding community structure and knowledge patterns.

Generally speaking, we can divide previous research of finding communities in networks into two main principle lines of research: *graph partitioning* and *hierarchical clustering*. These two lines of research are really addressing the same question, albeit by somewhat different means. There are, however, important differences between the goals of the two camps that make quite

different technical approaches desirable [27]. For example, *graph partitioning* approaches usually know in advance the number and size of the groups into which the network is to be split, while *hierarchical clustering* methods normally assume that the network of interests divide naturally into some subgroups, determined by the network itself and not by the user.

Graph Partitioning. There is a long tradition of research by computer scientists on graph partitioning [31]. Generally, finding an exact solution to a partitioning task is believed to be an NP-complete problem, making it prohibitively difficult to solve for large graphs. However, a wide variety of heuristic algorithms have been developed and give good solutions in many cases [12], e.g., multilevel partitioning [19], k-partite graph partitioning [20], relational clustering [21], flow-based methods [11], information-theoretic methods [8] and spectral clustering [30]. The main problem for these methods is that input parameters such as the number of the partitions and their sizes are usually required, but we do not typically know how many communities there are, and there is no reason that they should be roughly the same size. Various benefit functions have been proposed to avoid the problem, such as the *normalized cut* [33] and the *min-max cut* [9]. However, these approaches are biased in favour of divisions into equal-sized parts and thus still suffer from the same drawbacks that make graph partitioning inappropriate for community mining.

Hierarchical Clustering. The approaches developed by sociologists in their study of social networks for finding communities are perhaps better suited for our current purpose than the aforementioned clustering methods. The principle popular technique in use is *hierarchical clustering* [32]. The main idea of this technique is to discover natural divisions of social networks into groups, based on various metrics of similarity (usually represented as similarity x_{ij} between pairs (i, j) of vertices). The hierarchical clustering method has the advantage that it does not require the size or number of groups we want to find beforehand, therefore, it has been applied to various social networks with natural or predefined similarity metrics, such as the modularity and betweenness measure [6, 14, 25, 28]. However, they are usually slow and the performance depends highly on the corresponding metrics.

Recently, real world networks have been shown to have an overlapping community structure, which is hard to grasp with classical clustering methods where every vertex of the graph belongs to only one community. Based on these observations, fuzzy methods [15, 24, 29, 38] have been proposed for overlapping structure. Recent work by Xu et al. [37] proposed a fast SCAN algorithm to detect not only clusters, but also hubs and outliers in networks. However, the performance of these approaches depends on input parameters, which are very sensitive.

While all these methods successfully find communities, they implicitly assume that global information is always available. However, that is usually not the case for large networks in the real world. Clauset [5] and Luo et al. [22] proposed similar metrics for community detection with local informa-

tion, which are presented in detail in Section 3. Bagrow et al. proposed an alternative method to detect local communities [4], which spreads an l-shell outward from the starting node n, where l is the distance from n to all shell nodes. The performance of their approach depends on the parameter l and the starting node, because the result communities could be very different if the algorithm starts from border nodes instead of cores. The authors later proposed the "outwardness" metric Ω [3] to measure local structure, however, their method lacks an appropriate stopping criteria and thus still relies on arbitrary thresholds.

3 Preliminaries

Here we first define the problem of finding local communities in a network, then focus our efforts on reviewing existing algorithms.

3.1 Problem Definition

As mentioned in the introduction, local communities are densely-connected node sets that are discovered and evaluated based only on local information. Suppose that in an undirected network G (directed networks are typically first transformed to undirected ones), we start with perfect knowledge of the connectivity of some set of nodes, i.e., the known local portion of the graph, which we denote as D. (Note that D may start with one node, but can later contain a set of nodes and connections between them as a local community.) This necessarily implies that we also have limited information for another shell node set S, which contains nodes that are adjacent to nodes in D but do not belong to D (note "limited" means that the complete connectivity information of any node in S is unknown). In such circumstances, the only way to gain additional information about the network G is to visit some neighbour nodes s_i of D (where $s_i \in S$) and obtain a list of adjacent nodes of s_i. As a result, s_i is removed from S and becomes a member of D while additional nodes may be added to S as neighbours of s_i. This typical one-node-at-one-step discovery process for local community detection is analogous to the method that is used by web crawling systems to explore the WWW. Furthermore, we define two subsets of D: the core node set C, where any node $c_i \in C$ have no outward links, i.e., all neighbours of c_i belong to D; and the boundary node set B, where any node $b_i \in B$ has at least one neighbour in S. Figure 1 shows node sets D, S, C and B in a network. Similar problem settings can be found in [3, 4, 5, 22], however, the metrics used to discover and evaluate the local community are different, as explained in the next section.

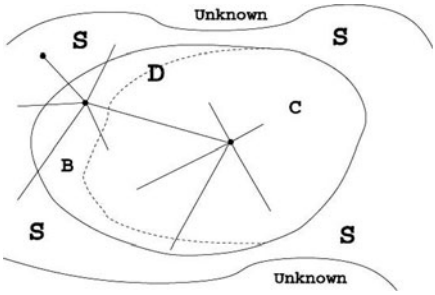


Fig. 1: Local Community Definition

3.2 Previous Approaches

Clauset has proposed the local modularity measure R [5] for the local community detection problem. R focuses on the boundary node set B to evaluate the quality of the discovered local community D.

$$R = \frac{B_{in\_edge}}{B_{out\_edge} + B_{in\_edge}} \tag{1}$$

where B_{in_edge} is the number of edges that connect boundary nodes and other nodes in D, while B_{out_edge} is the number of edges that connect boundary nodes and nodes in S. In other words, R measures the fraction of those "inside-community" edges in all edges with one or more endpoints in B. Therefore, the community D is measured by the "sharpness" of the boundary given by B.

Similarly, Luo et al. later proposed the measure called modularity M [22] for local community evaluation. Instead of measuring the internal edge fraction of boundary nodes, they directly compare the ratio of internal and external edges.

$$M = \frac{number\ of\ internal\ edges}{number\ of\ external\ edges} \tag{2}$$

where "internal" means two endpoints are both in D and "external" means only one of them belongs to D. An arbitrary threshold is set for M so that only node sets that have $M \geq 1$ are considered to be qualified local communities. M is strongly related to R. Consider a candidate node set D where every node in D has external neighbours, thus we have $|C| = 0$ and $B = D$, which means $B_{in\_edge} = internal\ edges$ and $B_{out\_edge} = external\ edges$. The threshold $M \geq 1$ is equivalent to $R \geq 0.5$. It is straight-forward to identify local communities with the R or M metric. Given a starting set D, in every step we merge the node into D from S which most increases the metric score, and then update D, B and S. This process is repeated until all nodes in S give negative value if merged in D, i.e., $\Delta R < 0$ or $\Delta M < 0$.

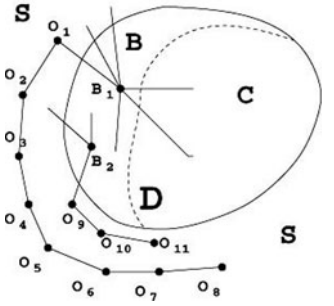


Fig. 2: Problem of Previous Approaches

Indeed algorithms using these metrics are able to detect communities in complex networks, however, their results usually include many outliers, i.e., the discovered communities have high recall but low accuracy, which reduces the overall community quality. Figure 2 illustrates the problem for R and M. In the figure, we have a local community D, its boundary B and nodes $O_1, ..., O_{11}$, which are outliers since they are barely related to nodes in D. Without loss of generality, let us assume that all nodes in S, except O_1 and O_9, will decrease the metric score if included in D. Now if we try to greedily maximize the metric R or M, all outliers (O_1 to O_8 and O_9 to O_{11}) will be merged into D, one by one. The reason is that every merge of node O_i does not affect the external edge number but will increase the internal edge number by one. Similarly, the algorithm would merge any node into D as long as it connects to the same number of nodes inside and outside the local community node set. Therefore, in addition to actual members, the resulting community would contain many weak-linked outliers, whose number can be huge for some networks, e.g., the WWW.

4 Our Approach

Existing approaches discussed in Section 3 are relatively simple: an effective local community detection method should be simple, not only because the accessible information of the network is restricted to merely a small portion of the whole graph, but also because the only means to incorporate more information about the structure is by expanding the community, by one node at one step. With these limitations in mind, we present our L metric and the local community discovery algorithm.

4.1 The Local Community Metric L

Intuitively, there are two factors one may consider to determine whether a node set in the network is a community or not: 1) high value node relations within the set, and 2) low value relations between inside nodes and the rest of the graph. Therefore, almost all existing metrics directly use the internal and external degrees to represent these two significant factors, and identify local communities by maximizing the former while minimizing the latter. However, their community results might include many outliers and the overall community quality is questionable (See Section 3.2 and Section 5.1.1 for examples). The important missing aspect in these metrics is the *connection density*, because is not the absolute number of connections that matters in community structure evaluation. For instance, even if there are one million edges within one node set N and no outward links at all, it is not sensible to identify N as a strong community if every node in N connects only one or two neighbours. We therefore propose to measure the community internal relation L_{in} by the average internal degree of nodes in D:

$$L_{in} = \frac{\sum_{i \in D} IK_i}{|D|} \tag{3}$$

where IK_i is the number of edges between node i and nodes in D. Similarly, we measure the community external relation L_{ex} by the average external degree of nodes in B:

$$L_{ex} = \frac{\sum_{j \in B} EK_j}{|B|} \tag{4}$$

where EK_j is the number of connections between node j and nodes in S. Note that L_{ex} only considers boundary nodes instead of the whole community D, i.e., the core nodes are not included since they do not contribute any outward connections. Now we want to maximize L_{in} and minimize L_{ex} at the same time. Fortunately, this can be achieved by maximizing the following ratio:

$$L = \frac{L_{in}}{L_{ex}} \tag{5}$$

Note that it is possible to quantify the density L_{ex} by other means, e.g., by using the average number of connections from the shell nodes to community nodes to measure L_{ex}. However, this method fails for the local community identification problem because the shell set is usually incomplete. For example, while the friend list of user A is available in Facebook, the list of the users that choose A as a friend is hard to obtain.

4.2 Local Community Structure Discovery

Using L to evaluate the community structure, one can identify a local community by greedily maximizing L and stopping when there are no remaining nodes in S that increases L if merged in D. However, this straight-forward method is not robust enough against outliers. Take Figure 2 as an example. Although L_{in} for O_1 would decrease because O_1 only connects to one node in D, the overall L might increase because the denominator L_{ex} decreases as well (O_1 only connects to one node outside D). Therefore, it is still possible to include outlier O_1 in the community. To deal with this problem, we look further into the metric instead of simply maximizing the score in a greedy manner. We note there are three situations in which we have an increasing L score. Assume i is the node in question and L'_{in}, L'_{ex} and L' are corresponding scores if we merge i into D, the three cases that will probably result in $L' > L$ are:

1. $L'_{in} > L_{in}$ and $L'_{ex} < L_{ex}$
2. $L'_{in} < L_{in}$ and $L'_{ex} < L_{ex}$
3. $L'_{in} > L_{in}$ and $L'_{ex} > L_{ex}$

Obviously nodes in the first case belong to the community since they strengthen the internal relation and weaken the external relation. Nodes in the second case, e.g., O_1 in Figure 2, are outliers. They are weakly connected to the community as well as the rest of the graph. Finally, the role of nodes in the third case cannot be decided yet, since they are strongly connected to both the community and the network outside the community. More specifically, when we meet a node i, which falls into this case during the local community discovery process, there are two possibilities. First, node i can be the first node of an enclosing community group that is going to be merged one by one; Second, i connects to many nodes, inside or outside the community, and can be referred to as a "hub." We do not want hubs in the local community. However, it is too early to judge whether the incoming node is a hub or not. Therefore, we temporarily merge nodes in the first and third cases into the community. After all qualified nodes are included, we re-examine each node by removing it from D and check the metric value change of its merge again. Now we only keep nodes in the first case. If node i is a member of an enclosing group, L'_{ex} should decrease because all its neighbours are now in the community as well, while hub nodes would still belong to the third case (See Algorithm 4). Finally, the starting node should still be found in D, otherwise, we believe a local community does not exist if we start from n_0. (See Algorithm 5.)

The computation of each L'_i can be done quickly using the following expression.

$$L'_i = \frac{\frac{Ind+2*Ind_i}{|D|+1}}{\frac{Outd-Ind_i+Outd_i}{|B'|}} \tag{6}$$

Algorithm 4 General Local Community Identification

Input: A social network G and a start node n_0.
Output: A local community with its quality score L.
1. Discovery Phase:
Add n_0 to D and B, add all n_0's neighbours to S.
do
for each $n_i \in S$ **do**
compute L'_i
end for
Find n_i with the maximum L'_i, breaking ties randomly
Add n_i to D if it belongs to the first or third case
Otherwise remove n_i from S.
Update B, S, C, L.
While $(L' > L)$
2. Examination Phase:
for each $n_i \in D$ **do**
Compute L'_i, keep n_i only when it is the first case
end for

Algorithm 5 Single Local Community Identification

Input: A social network G and a start node n_0.
Output: A local community D for node n_0.
1. Apply algorithm 4 to find a local community D for n_0.
2. If $n_0 \in D$, return D, otherwise there is no local community for n_0.

where Ind and $Outd$ are the number of within and outward edges of D before merging i, and should be updated after each merge; Ind_i and $Outd_i$ are the number of edges from node i to the community and the rest of network; B' is the new boundary set after examining all i's neighbour in D. In the discovery phase, L'_i need to be recomputed for every node in S to determine the one with the maximum ΔL, thus the complexity of the algorithm is $O(kd|S|)$, where k is the number of nodes in the D, and d is the mean degree of the graph. However, in networks for which local community algorithms are applied, e.g., the WWW, and where adding a new node to D requires the algorithm to obtain the link structure, the running time will be dominated by this time-consuming network information retrieval. Therefore, for real world problems the running time of our algorithm is linear in the size of the local community, i.e., O(k). Note that in Algorithm 4 we begin with only one node n_0, but the same process could apply for multiple nodes to allow a larger starting D, C, B and S.

4.3 Iterative Local Expansion

Algorithm 4 is for identifying one local community for a specific set of starting nodes. However, we could apply this algorithm iteratively to cover the whole graph or a large section of the graph if the iterative process is terminated. In other words, instead of one-node-at-one-step, we expand as one-community-at-one-step to discover the community structure in the network. See Algorithm 6.

Algorithm 6 Iterative Expansion Algorithm

Input: A social network G, a start node n_0 and the community number m (optional).
Output: A list of local communities.
1. Apply algorithm 4 to find a local community l_0 for n_0.
2. Insert neighbours of l_0 into the shell node set S
3. **While** $(|S|! = 0$ && $(i \leq m))$
   Randomly pick one node $n_i \in S$.
   Apply algorithm 4 to find a local community l_i
   for n_i.
   Remove n_i and nodes in S that are covered by l_i.
   Update S by neighbours of l_i that are not covered
   yet.
4. Output m local communities $l_0, l_1, l_2..., l_m$, m could be given as a stop parameter if necessary.

In algorithm 6, we recursively apply the local community identification algorithm to expand the community structure. Every time we find a local community, we update the shell node set, which is actually a set of nodes whose community information is still unclear. Note that here we accept identified local communities even if the starting node is not included. The shown algorithm stops when we have learned the whole structure of the network; however, we could also give parameters as stopping criteria if exploring the whole network is unnecessary or impractical, such as the number of discovered communities (m), or the number of nodes that has been visited (k). The algorithm could also be parallel and have multiple starting nodes, where several local community identification procedures start simultaneously from different locations of the network. Obviously, the complexity of the Algorithm 6 is still $O(kd|S|)$.

As previously discussed, in real world networks, one entity usually belongs to multiple communities. However, most of the existing approaches cannot identify such overlapping communities. Fortunately, even though we do not specifically focus on finding the overlapping property, our approach is able to discover overlapping communities, since in our algorithm nodes could be included in multiple local communities based on their connection structure.

5 Experiment Results

In this section we conducted several experiments to validate the effectiveness of the proposed approach.

5.1 Comparing Metric Accuracy

Since the ground truth of local communities in a large network is hard to define, previous research usually apply their algorithms on real networks and analyze the results based on common sense, e.g., visualizing the community structure or manually evaluating the relationship between disclosed entities [4, 5, 22]. Here we adapt a different method to evaluate the discovered local communities. We provide a social network with absolute community ground truth to the algorithm, but limit its access to network information to local nodes only. The only way for the algorithm to obtain more network knowledge is to expand the community, one node at a time. Therefore, we can evaluate the result by its accuracy, while satisfying limitations for local community identification. Based on our observations, the greedy algorithm based on metric R [5] (we refer to it as algorithm R) outperforms all other known methods for local community detection. Furthermore, similar to our approach, R does not require any initial parameters while other methods [3, 4, 22] rely on parameter selection. Therefore, in this section we compare the results of our algorithm and algorithm R on different real world networks to show that our metric L is an improvement for local community detection.

5.1.1 The NCAA Football Network

The first dataset we examine is the schedule for 787 games of the 2006 National Collegiate Athletic Association (NCAA) Football Bowl Subdivision (also known as Division 1-A) [37]. In the NCAA network, there are 115 universities divided into 11 conferences[1]. In addition, there are four independent schools, namely Navy, Army, Notre Dame and Temple, as well as 61 schools from lower divisions. Each school in a conference plays more often with schools in the same conference than schools outside. Independent schools do not belong to any conference and play with teams in all conferences, while lower division teams play only few games. In our network vocabulary, this network contains 180 vertices (115 nodes as 11 communities, 4 hubs and 61 outliers), connected by 787 edges.

[1] The ground truth of communities (conferences) can be found at http://sports.espn.go.com/ncf/standings?stat=index&year=2006

We provide this network as input to our algorithm and algorithm R. Every node in a community, which represents one of the 115 schools in an official conference, has been taken as the start node for both algorithms. Based on the ground truth posted online, the *precision*, *recall* and *f-measure* score, which is defined as the harmonic mean of precision and recall, of all the discovered local communities are calculated. We average the score for all schools in one conference to evaluate the accuracy of the algorithm to detect that particular community. Finally, an overall average score of the precision, recall and f-measure score of all communities is calculated for comparison.

The experiment results are shown in Table 1. We first note the disadvantage of metric R we reviewed theoretically in Section 3.2, which is vulnerability against outliers, has been confirmed by the results: for all communities, Algorithm R gets a higher recall but a much lower precision, which eventually leads to an unsatisfactory f-measure score. On the other hand, the accuracy of our algorithm is almost perfect, with a 0.952 f-measure score on average. Second, we see that our algorithm does not return local communities if starting with certain nodes in the network, namely 34 of the 115 schools representing 29.6%. (Note that in these cases the local community is considered not existent and is not included in the average accuracy calculation even though the starting nodes are not outliers.) However, this result actually shows merit of our approach instead of weak points. Generally speaking, in one local community, nodes can be classified into cores and peripheries. It would be easier for an algorithm to identify the local community if it began from cores rather than peripheries. For example, if the algorithm starts from a periphery node i in community c, the expansion step might fall into a different neighbour community d, which has some members connecting to i, due to lack of local information. It would be more and more difficult to return to c as the algorithm proceeds, because members of d are usually taken in one after another and finally, the discovered local community would be d plus node i, instead of c. Fortunately, our algorithm detects such phenomena in the examination phase since i will be found as an outlier to d. Therefore we do not return the result as a local community for i since we realize that it is misdirected in the beginning. As a possible solution for this problem, we can always start with multiple nodes, by which we provide more local information to avoid the possible misdirection. Note that while our algorithm handles such situations, algorithm R returns communities for every node without considering this problem, which is one reason for its low accuracy. Also note that another approach [22] has a similar "deletion step", however, that approach depends on arbitrarily selected thresholds.

5.1.2 The Amazon Co-purchase Network

While mid-size networks with ground truth provide a well-controlled testbed for evaluation, it is also desirable to test the performance of our algorithm on

2006 NCAA League		Algorithm Results						
		Algorithm using metric R			Algorithm using metric L			
Conference	Size	P	R	F	No Community	P	R	F
Mountain West	9	0.505	0.728	0.588	0 node	0.944	1	0.963
Mid-American	12	0.392	0.570	0.463	1 nodes	0.923	1	0.96
Southeastern	12	0.331	0.541	0.410	3 nodes	1	1	1
Sun Belt	8	0.544	0.891	0.675	3 nodes	1	1	1
Western Athletic	9	0.421	0.716	0.510	4 nodes	0.6	1	0.733
Pacific-10	10	0.714	1	0.833	0 nodes	1	1	1
Big Ten	11	0.55	1	0.710	9 nodes	0.729	1	0.814
Big East	8	0.414	0.781	0.534	5 nodes	1	1	1
Atlantic Coast	12	0.524	0.924	0.668	3 nodes	1	1	1
Conference USA	12	0.661	1	0.796	1 nodes	1	1	1
Big 12	12	0.317	0.465	0.355	5 nodes	1	1	1
Total	115	0.488	0.783	0.595	34 nodes (29.6%)	0.927	1	0.952

Table 1: Algorithm Accuracy Comparison for the NCAA Network (Precision (P), Recall (R) and F-measure (F) score are all average values for all nodes in the community).

Alg.	Items (Books) in the Local Community
Both	Smith of Wootton Major*
	LoR: A Reader's Companion#
	LoR: 50th Anniversary, One Vol. Edition*
	(The starting node) LoR [BOX SET]*
L	On Tolkien: Interviews, ... and Other Essays#
	Tolkien Studies: ... Scholarly Review, Vol. 2#
	Tolkien Studies: ... Scholarly Review, Vol. 1#
	... Grammar of an Elvish Language from LoR#
	J.R.R. Tolkien Companion and Guide#
	The Rise of Tolkienian Fantasy#
	... Celtic And Norse in Tolkien's Middle-Earth#
R	Farmer Giles of Ham & Other Stories*
	... Farmer Giles of Ham*
	Roverandom*
	Letters from Father Christmas, Revised Edition*
	Bilbo's Last Song*
	... Wonderful Adventures of Farmer Giles*
	Poems from The Hobbit*
	Father Christmas Letters Mini-Book*
	Tolkien: The Hobbit Calendar 2006*

Table 2: Algorithm Comparison for the Amazon Network. $*$ indicates the author is J.R.R. Tolkien while # is not.

	Items (Books) in the Local Communities
1	Mozart: A Cultural Biography
2	The Cambridge Companion to Mozart (Cambridge Companions to Music)
3	The Mozart Compendium: A Guide to Mozart's Life and Music
4	Mozart: The Golden Years
...	...
19	The Complete Mozart: A Guide to the Musical Works ...
1	Chopin In Paris: The Life And Times Of The Romantic Composer
2	The Cambridge Companion to Chopin (Cambridge Companions to Music)
3	Chopin (Master Musicians Series)
4	Chopin: The Man and His Music
5	Chopin's Letters
...	...
15	The Parisian Worlds of Frederic Chopin
1	The Cambridge Companion to Schubert (Cambridge Companions to Music)
2	The Cambridge Companion to Mozart (Cambridge Companions to Music)
3	The Cambridge Companion to Chopin (Cambridge Companions to Music)
4	The Cambridge Companion to Stravinsky (Cambridge Companions to Music)
5	The Cambridge Companion to Ravel (Cambridge Companions to Music)
...	...
9	The Cambridge Companion to Beethoven (Cambridge Companions to Music)
1	The New Webster's Grammar Guide
2	Hardcover, Longman Grammar of Spoken and Written English
3	Editorial Freelancing: A Practical Guide
4	The Oxford Dictionary for Writers and Editors
...	...
52	Modern American Usage: A Guide
1	Shakespeare's Language
2	Imagining Shakespeare
3	Hamlet: Poem Unlimited
4	... A Complete Pronunciation Dictionary for the Plays of William Shakespeare
...	...
66	William Shakespeare: A Compact Documentary Life

Table 3: Overlapping Local Community Examples for the Amazon Network

large real world networks. However, since ground truth of such large networks is elusive, we have to justify the results by common sense. We applied our algorithm and algorithm R to the recommendation network of Amazon.com, collected in January 2006 [22]. The nodes in the network are items such as books, CDs and DVDs sold on the website. Edges connect items that are frequently purchased together, as indicated by the "customers who bought this book also bought these items" feature on Amazon. Note that in this dataset we are looking for communities of "items" instead of communities of "people". There are 585,283 nodes and 3,448,754 undirected edges in this network with a mean degree of 5.89. Similar datasets have been used for testing in previous works [5, 22].

In table 2, we present discovered local communities for one popular book (*The Lord of the Rings (LOR)* by J.R.R. Tolkien), which is used as the start-

ing node. While both algorithms find communities, our algorithm detects books by authors other than Tolkien but are strongly related to the topic. On the other hand, more than 90% of the books in R's community are written by Tolkien. Moreover, after reading the reviews and descriptions on Amazon, we found that many of the books are for children, e.g, *Letters from Father Christmas*. These books are not related to dragons and magic, but are included in the community because they weakly connect to the starting node since they share the same author, as we discussed in Section 3.2.

5.2 Iteratively Finding Overlapping Communities

After evaluating the accuracy of the L metric and our algorithm for single community identification, here we apply Algorithm 6 on the Amazon network to find overlapping communities iteratively. Table 3 shows several local community examples of our result. Note that start nodes of some communities may be removed by our algorithm. Such communities are not included using Algorithm 5 for single local community identification in earlier experiments.

The first community has 19 nodes, originated at the book *Mozart: A Cultural Biography*. It naturally includes other books about the life and music of the legendary musician. Similarly, we have another 15-node-community about the famous Polish pianist Chopin. The third community is a book series, which is the *Cambridge Companions to Music*. Finally, the fourth community and fifth community contain books about English grammar and William Shakespeare. Note that many other global community detection algorithms, e.g., FastModularity [6], become slow for such huge networks. Moreover, they may not apply if the global network information is unavailable.

Aside from local communities of books in Amazon, our approach also finds overlaps between communities. For example, the two books *The Cambridge Companion to Mozart (Cambridge Companions to Music)* and *The Cambridge Companion to Chopin (Cambridge Companions to Music)* are found both in the community of the book series and the community of the subject. One could easily justify there is indeed some overlap.

6 Conclusion and Future Work

We have reviewed problems of existing methods for constructing local communities, and propose a new metric L to evaluate local community structure when the global information of the network is unavailable. Based on the metric, we develop a two-phase algorithm to identify the local community of a set of given starting nodes. Our method does not require arbitrary initial parameters, and it can detect whether a local community exists or not for

a particular node. Moreover, we extend the algorithm to an iterative local expansion approach to detect communities to cover large networks. We have tested our algorithm on real world networks and compared its performance with previous approaches. Experimental results confirm the accuracy and the effectiveness of our metric and algorithm.

In this work, we assume the social network to be "static". It would be interesting to investigate the possibility of extending the proposed metric and algorithms to discover communities in a dynamic social network. Our future work also includes the investigation of a means to validate the effectiveness of overlapping community detection in a large network without ground truth.

7 Acknowledgments

Our work is supported by the Canadian Natural Sciences and Engineering Research Council (NSERC), by the Alberta Ingenuity Centre for Machine Learning (AICML), and by the Alberta Informatics Circle of Research Excellence (iCORE). We wish to thank Eric Promislow for providing the Amazon data and Xiaowei Xu for the NCAA data.

References

1. L. A. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05*, pages 36–43, 2005.
2. S. O. Aral, J. P. Hughes, B. Stoner, W. Whittington, H. H. Handsfield, R. M. Anderson, and K. K. Holmes. Sexual mixing patterns in the spread of gonococcal and chlamydial infections. *American Journal of Public Health*, 89:825–833, 1999.
3. J. P. Bagrow. Evaluating local community methods in networks. *J.STAT.MECH.*, page P05001, 2008.
4. J. P. Bagrow and E. M. Bollt. Local method for detecting communities. *Physical Review E*, 72(4), 2005.
5. A. Clauset. Finding local community structure in networks. *Physical Review E*, 72:026132, 2005.
6. A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111, 2004.
7. L. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas. Comparing community structure identification. *J. Stat. Mech*, page P09008, 2005.
8. I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *KDD*, pages 89–98, 2003.
9. C. H. Q. Ding, X. He, H. Zha, M. Gu, and H. D. Simon. A min-max cut algorithm for graph partitioning and data clustering. In *ICDM*, pages 107–114, 2001.
10. J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Phys. Rev. E*, 72:027104, 2005.
11. G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *KDD*, pages 150–160, 2000.
12. S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

13. G. P. Garnett, J. P. Hughes, R. M. Anderson, B. P. Stoner, S. O. Aral, W. L. Whittington, H. H. Handsfield, and K. K. Holmes. Sexual mixing patterns of patients attending sexually transmitted diseases clinics. *Sexually Transmitted Diseases*, 23:248–257, 1996.
14. M. Girvan and M. Newman. Community structure in social and biological networks. In *PNAS USA, 99:8271-8276*, 2002.
15. S. Gregory. An algorithm to find overlapping community structure in networks. In *PKDD*, pages 91–102, 2007.
16. R. Guimera and L. A. N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, 2005.
17. S. Gupta, R. M. Anderson, and R. M. May. Networks of sexual contacts: Implications for the pattern of spread of hiv. *AIDS*, 3:807–817, 1989.
18. D. Jensen. Statistical challenges to inductive inference in linked data, 1999.
19. G. Karypis and V. Kumar. Multilevel k-way partitioning scheme for irregular graphs. *Journal of Parallel and Distriuted Computing*, 48(1):96–129, 1998.
20. B. Long, X. Wu, Z. M. Zhang, and P. S. Yu. Unsupervised learning on k-partite graphs. In *KDD*, pages 317–326, 2006.
21. B. Long, Z. M. Zhang, and P. S. Yu. A probabilistic framework for relational clustering. In *KDD*, pages 470–479, 2007.
22. F. Luo, J. Z. Wang, and E. Promislow. Exploring local community structures in large networks. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 233–239, 2006.
23. M. A. Nascimento, Jörg Sander, and J. Pound. Analysis of sigmod's co-authorship graph. *SIGMOD Record*, 32(2):57–58, 2003.
24. T. Nepusz, A. Petroczi, L. Negyessy, and F. Bazso. Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E*, 77, 2008.
25. M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, 2004.
26. M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74, 2006.
27. M. E. J. Newman. Modularity and community structure in networks. *PNAS USA*, 103, 2006.
28. M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69, 2004.
29. G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814, 2005.
30. J. Ruan and W. Zhang. An efficient spectral algorithm for network community discovery and its applications to biological and social networks. In *ICDM*, pages 643–648, 2007.
31. S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1:27–64, 2007.
32. J. Scott. Social network analysis: A handbook, Sage, London 2nd edition(2000).
33. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE. Trans. on Pattern Analysis and Machine Intelligence*, 2000.
34. A. F. Smeaton, G. Keogh, C. Gurrin, K. McDonald, and T. Sodring. Analysis of papers from twenty-five years of sigir conferences: What have we been doing for the last quarter of a century. *SIGIR Forum*, 36(2):39–43, 2002.
35. J. R. Tyler, D. M. Wilkinson, and B. A. Huberman. Email as spectroscopy: automated discovery of community structure within organizations. *Communities and technologies*, pages 81–96, 2003.
36. S. White and P. Smyth. A spectral clustering approach to finding communities in graphs. In *SIAM*, 2005.
37. X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger. Scan: a structural clustering algorithm for networks. In *KDD*, pages 824–833, 2007.
38. S. Zhang, R. Wang, and X. Zhang. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A*, 374:483–490, 2007.

Why Do Diffusion Data Not Fit the Logistic Model? A Note on Network Discreteness, Heterogeneity and Anisotropy

Dominique Raynaud

Abstract Diffusion of innovations and knowledge is in most cases accounted for by the logistic model. Fieldwork research however constantly report that empirical data utterly deviate from this mathematical function. This chapter scrutinizes network forcing of diffusion process. The departure of empirical data from the logistic function is explained by social network discreteness, heterogeneity and anisotropy. New indices are proposed. Results are illustrated by empirical data from an original study of knowledge diffusion in the medieval academic network.

Introduction

Diffusion of innovations is the process by which an innovation is communicated through certain channels over time among the members of a social system [27, p. 5]. Diffusion of knowledge may be defined in the same manner, replacing what should be in the previous definition. Social network analysis is in its early stages of application to diffusion issues.

> Compared with other aspects of diffusion research, there have been relatively few studies of how the social or communication structure affects the diffusion and adoption of innovations in a system [27, p. 25].

So speaks Everett M. Rogers, the outstanding promoter of diffusion studies, about the way they are connected to network analysis. In the 1970's, according to a content analysis of 1,084 empirical publications, diffusion networks represented less than one percent of diffusion research. Ten years ago, a book especially dedicated to network analysis has the same diagnosis:

Dominique Raynaud
PLC, Université de Grenoble, F-38040 Grenoble Cedex 9, and GEMASS, CNRS UMR 8598 / Université Paris Sorbonne, 54 bd Raspail, F-75000 Paris, France
e-mail: dominique.raynaud@upmf-grenoble.fr

> Only few material coupling a diffusion study with network analysis is available [8, p. 189].

The authors consequently content themselves with classical studies in the field. The fact that different authors interested in the diffusion of innovations vs. network analysis have detected the same lack of research, vouches for a promising fieldwork.

1 A Brief Historical Sketch

Despite the fact that the first paper explicitly dedicated to network diffusion was written in 1979 by Everett M. Rogers [26], the concern is much more ancient. For instance, in the course of his studies on the cholera, in 1884, Etienne-Jules Marey already applied network perspective to diffusion data. He had the insight that the topology of social network determines the form of the diffusion process. Marey says: "Closed institutions: prisons, boarding schools, convents, asylums, etc., usually escape to cholera; but if it gets into, it takes a terrible toll of victims" [20, p. 670]. Cliques (i.e. closed communities) exhibit atypical behaviour: they are resistant to the disease or completely devastated by the epidemic. The "clique effect" was independently rediscovered in 1973 by Mark S. Granovetter [12]: "weak ties" favour diffusion, "strong ties" protect the members from a tentative adoption: either they all adopt, or they all reject.

The first book approaching the diffusion of innovations through network analysis came out in 1995 by a student of Rogers: Thomas W. Valente [35]. His scope was to compare three classical datasets: the adoption of tetracycline by 130 physicians of the Middle West [6]; the diffusion of innovations among 692 Brazilian farmers [28]; the diffusion of family planning in 24 Korean villages [29]. Held with fifteen years of hindsight, the book appears a little disapointing for seven chapters out of nine are in fact dedicated to apply thresholds and critical mass models to diffusion issues. Network analysis occupies thirty pages [35, p. 31-61], where classical measures of density, centrality and equivalence are processed out on the three datasets.

In the past decades, diffusion process have been simulated either within deterministic or probabilistic models [1, 19, 16, 5, 9, 17]. Researchers have explored a full range of simulations, including Monte Carlo, Ising, Potts, Krause-Hegselmann and Deffuant models [16, 10, 32, 34, 4]. However, more often than not, models presuppose the population to be homogeneous. Simulations are implemented on regular bi-dimensional lattices. Modeling rarely assumes the topology in which the diffusion occurs, and it is only in the 2000s that sociologists, economists and physicists addressed the point [7, 15, 30, 31, 18, 2]. The network-based approach of diffusion is full of consequences, some of which are still developing.

There is consequently little research on the irregularity of diffusion curves. The state-of-the-art is the following: 1/ Threshold models have relied the γ curve slope to higher or lower individual adoption thresholds [13, 14, 36]. 2/ Critical mass models have shown that, the more the early majority members have a high centrality index, the faster the critical mass is reached, and the higher is the saturation threshold [23, 35]. 3/ In a forthcoming paper, we find, with no supplementary details, that "important consequences of this large variability [of human behaviour] are the slowing down or speed up of information" [18, p. 6].

Diffusion irregularity is not the main concern of papers in statistical physics that closely combine network analysis with diffusion processes. They consider instead: 1/ critical points for avalanches [15]; 3/ random walks, qua they provide network topology characteristics [30]; 4/ modularity-based partitioning of graphs [31]; 5/ mean-field theory applications [2]. An overview of the current research would say that diffusion is basically seen as a *means* to find network properties. Hence, the literature is still lacking for network-based studies that could answer the long-delayed question of diffusion anomalies.

2 The Logistic Function

The choice of the mathematical model accounting for diffusion depends on the empirical conditions of spreading. If the empirical process is a step-by-step diffusion through interpersonal contacts, it follows the logistic function, which was first discovered in 1838 by Pierre-François Verhulst—albeit he never used the term [38]. In the recent years, there was a propensity to emancipate from his formalism; see [5]. In the following equation, the term $e^{-\gamma t}$ is the most characteristic of the process:

$$n_t = \frac{N'}{1 + ae^{-\gamma t}} \tag{1}$$

n_t represents the cumulative number of individuals having adopted at time t; N' the number of susceptible adopters in the given population; a is a parameter setting the number of early adopters; γ the slope of the curve at the inflection point (when $\gamma \to 0$, the curve is flat).

The bell-shaped first derivative represents the instant number of adopters at time t. The second derivative is swing-shaped. First and second derivatives respond to equations (2) and (3) respectively:

$$\frac{dn}{dt} = \frac{\gamma N' a e^{-\gamma t}}{1 + ae^{-\gamma t}} \tag{2}$$

$$\frac{d^2n}{dt^2} = \frac{\gamma^2 N' a e^{-\gamma t}(a e^{-\gamma t} - 1)}{(1 + a e^{-\gamma t})^3} \tag{3}$$

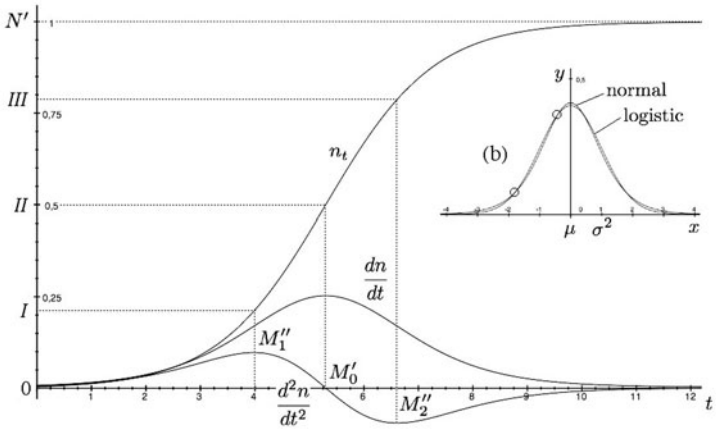


Fig. 1: The logistic function and its derivatives

This approach provides a natural classification of adopters through diffusion-based concepts. Introducing N' in the equations, we find that—whatever be parameters a and γ—early adopters, early majority, late majority and laggards appear at exactly 0.21, 0.5 and 0.79 of the susceptible population. These values are somewhat different from those given in Rogers' influential book [27], and afterwards endlessly repeated in the literature. According to him, early adopters, early majority, late majority and laggards appear respectively at 0.16, 0.5 and 0.84 of the susceptible population.

The reason for such a disagreement is that the classification of adopters is envisaged through common statistical concepts. The logistic first derivative is assimilated to a normal distribution. Rogers' values are in fact those of the mean μ and $\mu - \sigma^2$, $\mu + \sigma^2$. Valente is even more explicit:

> The logistic function has an inflexion point at 50 % adoption and two second-order inflection points, one each at one standard deviation below and above the mean [35, p. 83].

Standard deviation concept is inappropriate in this context: logistic function is not a Gaussian distribution. There is a simple visual proof for that: respective curves cross several times with each other, as in Fig. 1 (b). As a consequence, the values 0.16, 0.5 and 0.84 must be discarded and replaced by 0.21, 0.5 and 0.79.

3 Empirical Data

Does the logistic function fit the diffusion data? Compared to fieldwork data, logistic law rather appears as a mathematical ideality. In fact, curves never exhibit so a smooth and regular profile. Let us begin by the fact fieldwork studies constantly report that curves of diffusion are affected by irregularities. One can find in Valente's valuable book [35] a report on such irregularities (Fig. 2).

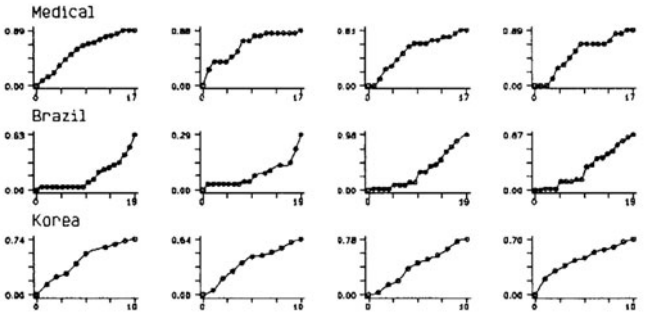


Fig. 2: Diffusion curves irregularity.

Such anomalies are evidently not accountable within the logistic model. We owe to Rapoport and Yuan the first penetrating remarks. While reaching the conclusion that "actual processes can be approximated by equivalent epidemics in well-mixed populations or by equivalent random social nets," they are nevertheless warning us that this conjecture "does not apply to highly structured populations" [24, p. 344-345]. The gap between the model and the real world lies in the assumption that societies are "well-mixed populations," thus assimilating the adoption of innovation to a random draw. Social network heterogeneity urges to abandon this assumption, which is the basis of all SIR—susceptible/infective/removed—epidemiological models.

During the 2000s, statistical physicists took part in the debate, by criticizing the logistic model as "utterly inaccurate" [22]. They have again argued that standard epidemiological models do not take into account that diffusions happen in highly structured, heterogeneous population, whose topology influences the very form of the diffusion [39, 21, 22]. This observation put the answer at hand.

4 Accounting for Anomalies

Let us extract empirical data from a recent study of knowledge diffusion within the academic network of medieval universities (XIIIth-XIVth centuries). In this special case, the social network is represented by a deterministic articulated k^+-regular graph [25]. This graph $G = (V, E)$ is particularly fitted for the study of diffusion anomalies, because it is little dense: $\delta = \ell / (g\,(g-1)) = 0.069$ and little cohesive: $\chi = (2 \sum_{i=1}^{g} \sum_{j=1}^{g} x_{ij} \;/\; (v_i, v_j) \Rightarrow (v_j, v_i)) / (g(g-1)) = 0.047$. In the following sections, these data will be used to illustrate theoretical findings.

The deviation of diffusion curves from the logistic model appears by superimposing the curves that have any vertex of the graph as emitter (Fig. 3). Since the set is heterogeneous, we may decompose it into subsets. Let us for instance isolate all curves whose receptor belongs to component $\mathcal{P}_7$ (Fig. 4).

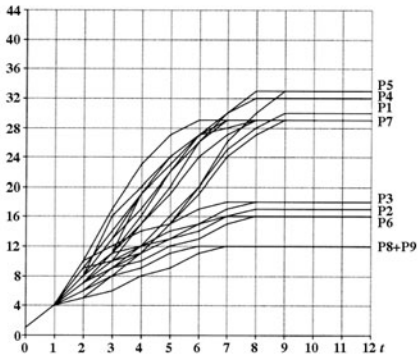


Fig. 3: Diffusion (any receptor)

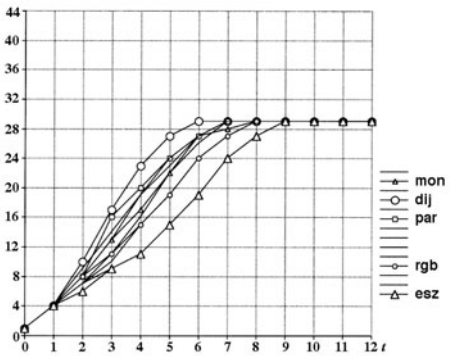


Fig. 4: Diffusion (receptor $\in \mathcal{P}_7$)

We are now in position to investigate the main causes why empirical diffusion curves deviate from the logistic function. Let us call, by license, "fast diffusion" any diffusion that converts many vertices in few steps; "slow diffusion" the one that eiher converts few vertices, or requires many steps to convert them all. Analytical concepts presented hereinafter are devised to fit diffusion networks that are always directed graphs and, more often than not, *grounded networks*—i.e. dependent on the position of vertices in space.

5 Net Discreteness

Discreteness of social networks results in slowing down/accelerating adoption, hence flattening/straightening the diffusion curve.

The more obvious difference between the logistic function (Fig. 1) and empirical curves (Fig. 3 and 4), is that the former is a smooth curve, when the latter is angled. While the logistic function is defined on a continuous (infinite) set, social diffusion happens within discrete (finite) structures. The lower the number of susceptible adopters, the more angled the curve.

There is a simple estimate for network discreteness. Let us call v_i the emitter vertex of any given diffusion process. The number of diffusion steps will be equal to the lenght of the longest geodesic d_{ik} originating in v_i. Local discreteness ranges from 0 when $d_{ik} = \infty$, to 1 when $d_{ik} = 1$, hence:

$$D_i = \frac{1}{\max(d_{ik})} \qquad D_i \in [0, 1] \tag{4}$$

If we need a comparative index, we first calculate the mean discreteness D_G—where g is the total number of vertices—, then the difference $D'_i = D_i - D_G$:

$$D'_i = \frac{1}{\max(d_{ik})} - \frac{1}{g}\sum_{j=1}^{g}\frac{1}{\max(d_{jk})} \qquad D'_i \in [-1, +1] \tag{5}$$

Local discreteness is null if the geodesic on which the vertex stands is as long as the graph mean geodesics. Positive values occur if $D_i \geqslant D_G$, when the region in which the information is going through is less dense than the rest of the graph, thus slowing down the diffusion; negative values otherwise. Discreteness values on the academic network are given in Appendix (columns D_i and D'_i). The data (ROS, 0.111, −0.048) are to be read as: "All vertices standing on the longest geodesic attached to ROS, i.e. {ROS, KOL, STR, DIJ, AST, GEN, PIS, SIE, PER, TOD, ROM} have discreteness 0.111 or, equivalently, are 0.048 less discrete than the rest of the graph."

6 Net Heterogeneity

Heterogeneity of social networks results in slowing down/accelerating adoption, hence flattening/straightening the diffusion curve.

Let us consider anew all curves of diffusion (Fig. 4). At $t = 2$, the curves, which were hitherto undifferentiated, split into five distinct profiles. The message from DIJ is adopted by six new vertices (fast growth); the one from ESZ is adopted by only two new vertices (slow growth). This difference is due to the fact the first message is diffused within two distinct subsets, when the second message never leaves the starting subset, composed of few susceptible adopters. The diffusion growth is as fast as the conversion rate. It is only when the message gets into new components of the network that it can cause mass adoption. Consequently, slow and fast phases are sensitive to articulators.

Network forcing is detectable by visual inspection. Let us consider STR vs. FLO as emitters. Two steps after he has quitted STR, the message propagates at constant rate $e_{i_t} = 5$ (Fig. 5). In the meanwhile, after having been constant until $t = 3$, FLO message adoption rate decreases to $e_{i_t} = 0$ (Fig. 6). Vertex STR connects immediately to two components, when FLO can join few vertices only. As a result, at $t = 5$, the message has touched 24 vertices (STR) vs. 12 vertices (FLO). Network properties subtending this contrastive behaviour are: STR articulator status and low transitivity ($T = 0.166$) vs. FLO confinement and high transitivity ($T = 0.5$), that is, high vs. low area heterogeneity.

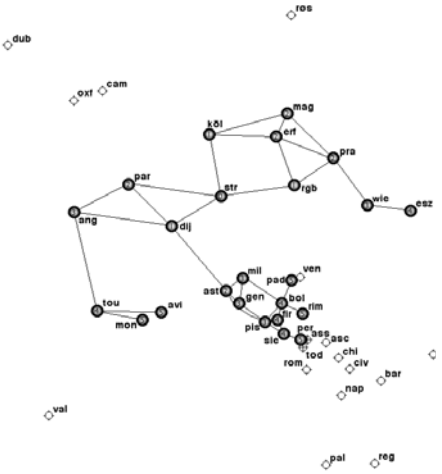


Fig. 5: Fast diffusion from STR

Fig. 6: Slow diffusion from FLO

The concept of social heterogeneity has become familiar since the discovery of "structural holes" by Peter Blau [3]. Several methods exist to detect structural holes and heterogeneity in a graph [11, 33]. We can proceed with by defining local density for vertex v_i—thus adapting graph density $\delta_G = L/g(g-1)$ to a local neighbourhood. Let us first define the neighbourhood as the circle of center v_i and radius $r = \max(\mathrm{d}_{ij})$, equal to the Euclidean distance of the remotest vertex v_j to which v_i is connected (whatever be the direction of the arc). Let ℓ_i be the outdegree of vertex v_i, and k_i the number of possible arcs capable of being built in this region. Local heterogeneity is:

$$H_i = 1 - \frac{\ell_i}{k_i} \qquad H_i \in [0, 1] \tag{6}$$

If we prefer a comparative index, we first calculate the mean heterogeneity H_G on the graph, then the difference $H'_i = H_i - H_G$:

$$H'_i = \left(1 - \frac{\ell_i}{k_i}\right) - \frac{1}{g}\sum_{j=1}^{g}\left(1 - \frac{\ell_j}{k_j}\right) \text{ that is equal to :}$$

$$H'_i = \frac{1}{g}\sum_{j=1}^{g}\left(\frac{\ell_j}{k_j}\right) - \frac{\ell_i}{k_i} \qquad H'_i \in [-1, +1] \tag{7}$$

This index is null when the region the information is going through is as dense as the whole graph. Positive values occur when $H_i \geqslant H_G$, that is when the local region is more heterogeneous than the rest of the graph, thus slowing down the diffusion; negative values otherwise. See Appendix for academic network values (columns H_i and H'_i). For example, the data (ROS, 0.000, -0.323) are to be read as: "Vertex ROS has null heterogeneity or, equivalently, is 0.323 less heterogeneous than the rest of the graph."

7 Net Anisotropy

Anisotropy of social networks results in slowing down/accelerating adoption, hence flattening/straightening the diffusion curve.

Let us isolate curves originating in DIJ and ESZ, two vertices having the greatest difference among component $\mathcal{P}_7$. A good way to link irregularity of curves with network properties is to represent the message progression within the network by joining all vertices converted at the same time by "isochrones." Fast spreading is associated to even isochrones when slow diffusion is coupled with uneven isochrones. As a result, six to nine steps are needed so that DIJ and ESZ information could cover the same area (Fig. 7 and 8). At any first steps, DIJ can distribute the message all around (quasi-isotropy), when ESZ can send information to its western neighbours only (anisotropy). This is a last factor explaining anomalies.

As far as I know, network anisotropy has never been defined in network analysis literature. This concept could be derived from other fields of physics, as optics, electricity or magnetism, but in the special case of grounded social networks, there is a more direct method to define the concept of anisotropy. Suppose vertex v_i has an outdegree ℓ_i and let α_{mn} be the angle of two adjacent arcs. The isotropic case arises when every angle is equal to $2\pi/\ell_i$. In order to appreciate anisotropy, we define the deviation of angle α_{mn} from $2\pi/\ell_i$. Vertex local anisotropy will then be the sum of actual deviations divided by the sum of theoretical maximum deviations:

$$A_i = \frac{\left|\alpha_{mn} - \frac{2\pi}{\ell_i}\right| + \left|\alpha_{no} - \frac{2\pi}{\ell_i}\right| + \ldots + \left|\alpha_{sm} - \frac{2\pi}{\ell_i}\right|}{2\,(\ell_i - 1)\frac{2\pi}{\ell_i}}$$

Since angles $\alpha_{mn}, \alpha_{no} \ldots$ are supplementary, we may simplify:

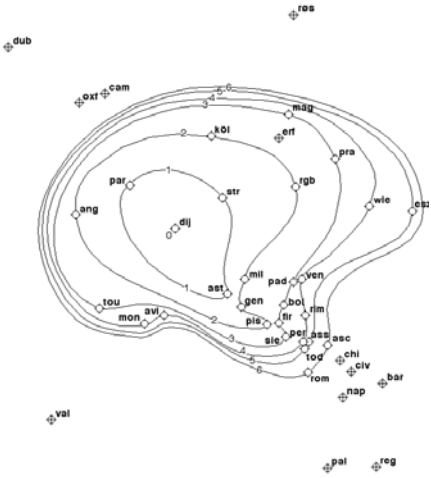


Fig. 7: Fast diffusion from DIJ

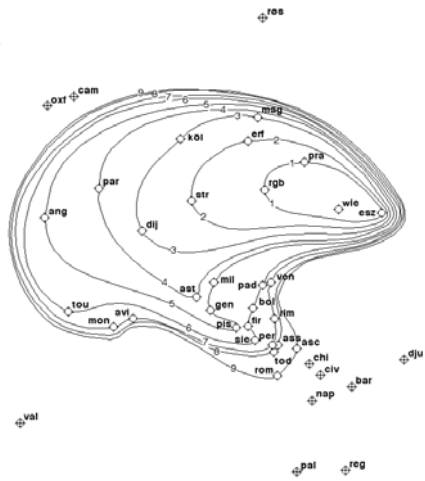


Fig. 8: Slow diffusion from ESZ

$$A_i = \frac{2\,\max\left|\alpha_{mn} - \dfrac{2\pi}{\ell_i}\right|}{2\,(\ell_i - 1)\dfrac{2\pi}{\ell_i}} \therefore$$

$$A_i = \frac{\ell_i\ \max\left|\alpha_{mn} - \dfrac{2\pi}{\ell_i}\right|}{2\pi(\ell_i - 1)} \qquad A_i \in [0, +1] \tag{8}$$

The comparative index is obtained as previously, by calculating the mean anisotropy A_G on the graph (with g vertices), then the difference $A'_i = A_i - A_G$:

$$A'_i = \frac{1}{2\pi}\left[\frac{\ell_i\ \max\left|\alpha_{mn} - \dfrac{2\pi}{\ell_i}\right|}{(\ell_i - 1)} - \frac{1}{g}\sum_{j=1}^{g}\frac{\ell_i \max\left|\alpha_{pq} - \dfrac{2\pi}{\ell_j}\right|}{(\ell_j - 1)}\right] \quad A'_i \in [-1, +1] \tag{9}$$

Local anisotropy is null when the local region has the same degree of isotropy as the rest of the graph. Positive values occur in the case $A_i \geqslant A_G$, when the region the message is going through is less isotropic than the rest of the graph, thus slowing down the diffusion; negative values otherwise. See Appendix for academic network data (columns A_i and A'_i). The data (ROS, 0.866, +0.388) are to be read as: "Vertex ROS has anisotropy 0.866 or, equivalently, is 0.388 more anisotropic than the rest of the graph."

8 A Test of DHA Indices

First of all, the definition of D (discreteness), H (heterogeneity) and A (anisotropy) enables us to tell the difference between regular and irregular graphs. In a 2D square lattice, all geodesics are infinite, the term $1/d_{ik}$ is null and $D = 0$. Since all vertices are disposed regularly, and have the same outdegree, the term $\ell_i/k_i = +1$ everywhere, thus $H = 0$. Finally, since angles of all adjacent arcs are equal, any term $\alpha_{mn} = 2\pi/\ell_i$ so that $A = 0$. We recognize a well known property: a lattice is a perfect infinite, homogeneous and isotropic network.

Social networks utterly differ from this model. Regarding the conditions of spreading in the same graph, DHA indices make it possible to see the regions where the diffusion is speeding up, and where it is slowing down. Let us resume the discussion of the medieval academic network and compare the indices of all vertices (see Appendix). Network mean values are: discreteness $D_G = 0.159$, heterogeneity $H_G = 0.323$, anisotropy $A_G = 0.478$. Because of the problem of sign combination, it is not appropriate to agregate directly DHA indices. General results are nonetheless accessible.

Maximum discreteness occurs for vertices in the graph innermost component {AST, MIL, PAD, VEN, GEN, BOL, PIS, FIR, SIE, RIM, PER} because several directed filters impede information to propagate outside the component, thereby reducing the number of steps of diffusion. Maximum heterogeneity occurs for vertices situated on the borders of the central component {AST, PIS, ROM, CHI, CIV} because those vertices are both in dense regions and in position to be tied to many vertices with which they have no actual relation. Maximum anisotropy occurs for the outermost vertices of the graph (DUB, ROS, ESZ, DJU) because they can send information only inwards the network.

All vertices are now sorted into eight classes, according to the sign of their indices (see Appendix, last column):

$---$	{OXF, CAM, WIE}
$+--$	{MON, AVI, GEN, FIR, PAD}
$-+-$	{PRA, ERF, RGB, STR, PAR, VAL, ASC, CHI, CIV}
$--+$	{LIS, SEV, SAL, TOU, ANG, DUB, ROS, ESZ, DJU, REG, PAL}
$++-$	{DIJ, BOL, SIE, PER, ASS, TOD}
$+-+$	{MIL, VEN}
$-++$	{MAG, KOL, BAR, NAP}
$+++$	{AST, PIS, ROM, RIM}

The behaviour of classes $(---)$ and $(+++)$ is quite obvious. Other classes can either accelerate or slow down the diffusion.

A 3D-plotting ($x = D, y = H, z = A$) shows the wide dispersion of vertices on both HA axes (Fig. 9), while D seems to have a much more limited impact on how they are spatially distributed (Fig. 10). This limitation is not a general property. It is due to the academic network small diameter, $\phi = 9$. I have

plotted in light blue low-valued vertices, for which the corresponding vector has lenght $x \leqslant 0.300$ {MAG, ERF, PRA, TOU, VAL, BOL, SIE, PER, TOD, ASC}.

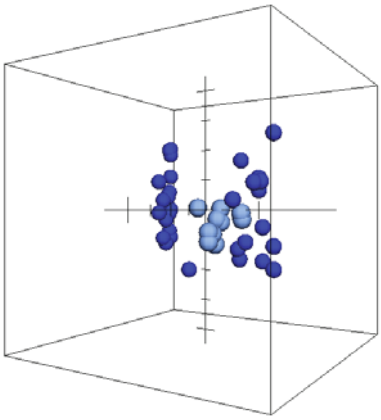


Fig. 9: DHA dispersion of vertices

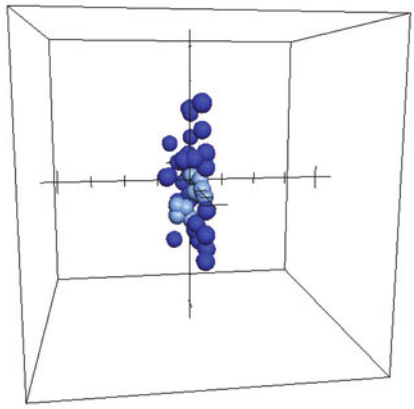


Fig. 10: D index limited impact

In DHA space, vertices that are speeding up vs. slowing down diffusion are separated by the plane Γ of equation:

$$-x - y - z = 0 \tag{10}$$

Accordingly, the impact of each node on the diffusion shape can be estimated by the Euclidean distance of any vertex v_i of coordinates (D'_i, H'_i, A'_i) to the plane Γ:

$$\omega_i = \frac{1}{\sqrt{3}} \left| - D'_i - H'_i - A'_i \right| \tag{11}$$

As a result, this compounded index (Appendix, column ω_i) tells us how much the network vertex is contributing to the deviation from the logistic curve. When ω_i sign is negative (yellow) the diffusion is accelerating, when it is negative (blue) the diffusion is slowing down (Fig. 11).

Focus now on the most contrastive classes of network vertices. Nodes which are speeding up the diffusion are in general situated in the innermost parts of the outer components {WIE, CAM, OXF, AVI, MON, SAL, SEV}. Conversely, nodes which are slowing down the diffusion are to be found in the outermost parts of the inner components {AST, PIS, RIM, CHI, ROM}; while out-out vertices {LIS, TOU, ANG, DUB, ROS, ESZ, MIL, PAD, VEN, DJU, REG, PAL} and in-in vertices {SIE, PER, ASS, TOD} have actually little impact on the diffusion.

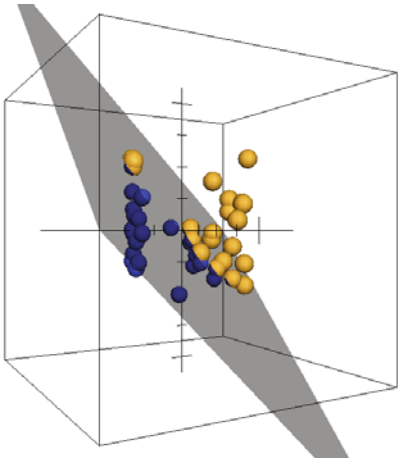


Fig. 11: DHA values

9 Conclusion

We are now in position to answer the issue adressed at the beginning of the chapter: *Diffusion data do not fit the logistic model because social space is discrete, heterogeneous and anisotropic*. Social space is thus an inadequate metaphor that should be outlawed. Any social spreading of innovations and knowledge is network-dependent. The agenda for future research should plan to: 1/ test DHA indices robustness—especially the concept of "net anisotropy" presented in these pages, 2/ focus on how DHA indices could predict the precise shape of a diffusion process occuring in a given social network and, conversely, 3/ deduce, from any empirical data on social diffusion, the DHA indices of the social network in which the diffusion takes place.

References

1. Bailey, N.T.J.: The Mathematical Theory of Epidemics. Charles Griffin, London (1957)
2. Baronchelli, A., Pastor-Satorras, R.: Diffusive dynamics on weighted networks. Preprint: arXiv: 0907.3810v1 (2009) [cond-mat.stat.mech]
3. Blau, P.: Inequality and Heterogeneity. The Free Press, New York (1977)
4. Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. Rev. Mod. Phys. **81** 591 (2009). Preprint: arXiv:0710.3256v1 [physics.soc-ph]
5. Cavalli-Sforza, L.L., Feldman, M.W.: Cultural Transmission and Evolution: A Quantitative Approach. Princeton University Press, Princeton (1981)
6. Coleman, J.S., Katz, E., Menzel, H.: The diffusion of an innovation among physicians. Sociometry **20** 253–270 (1957)
7. Cowan, R., Jonard, N.: Network Structure and the Diffusion of Knowledge. MERIT Technical report 99028, Maastricht University (1999)

8. Degenne, A., Forsé, M.: Introducing Social Networks. Sage Publications, London (1999)
9. Doran, J.E. and Gilbert, G.N. (eds.) Simulating Societies: The Computer Simulation of Social Phenomena. UCL Press, London (1994)
10. Doreian, P.: Mapping networks through time. In: Weesie, J. and Flap, H. (eds.) Social Networks through Time, pp. 245–264. ISOR/University of Utrecht (1990)
11. Fararo, T.J.: Biased networks and social structure theorems. Social Networks **3** 137–159 (1981)
12. Granovetter, M.S.: The strength of weak ties. Am. Journal Soc. **78** 1360–1380 (1973)
13. Granovetter, M.S.: Threshold models of collective behavior. Am. Journal Soc. **83** 1420–1443 (1978)
14. Granovetter, M.S., Soong, R.: Threshold models of diffusion and collective behavior. Journal Math. Soc. **9** 165–179 (1983)
15. Guardiola, X., Diaz-Guilera, A., Prez, C.J., Arenas, A., Llas, M.: Modelling diffusion of innovations in a social network. Phys. Rev. E **66** 0206121 (2002)
16. Hägerstrand, T.: A Monte Carlo approach to diffusion. Eur. Journal Soc. **6** 43–67 (1965)
17. Hegselmann, R., Mueller, U., Troizsch, K.G. (eds.) Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View. Kluwer Academic Publishers, Dordrecht (1996)
18. Iribarren, J.L., Moro, E.: Information diffusion epidemics in social networks. Preprint: arXiv: 0706.0641v1 (2008) [physics.soc-ph]
19. Kendall, D.G.: Mathematical Models of the Spread of Infection. Medical Research Council, London (1965)
20. Marey, E.J.: Les eaux contaminées et le choléra. CRAS **99** 667–683 (1884)
21. Moreno, Y., Pastor-Satorras, R., Vespignani, A.: Epidemic outbreaks in complex heterogeneous networks. Eur. Phys. Journal B **26** 521–529 (2002)
22. Newman, M.E.J.: The spread of epidemic disease on networks. Phys. Rev. E **66** 016128 (2002)
23. Oliver, P.E., Marwell, G., Teixera, R.: A theory of critical mass, I. Interdependance, group heterogeneity, and the production of collective action. Am. Journal Soc. **91** 522–556 (1985)
24. Rapoport, A., Yuan, Y.: Some experimental aspects of epidemies and social nets. In: Kochen, M. (ed.) The Small World, pp. 327–348. Ablex Publishing Company, Norwood (1989)
25. Raynaud, D.: Etudes d'épistémologie et de sociologie des sciences, 2. Pourquoi la perspective a-t-elle été inventée en Italie centrale? Habilitation à diriger les recherches. Université de Paris Sorbonne, Paris (2004)
26. Rogers, E.M.: Network analysis of the diffusion of innovations. In: Holland, P.W. and Leinhardt, S. (eds.) Perspectives on Social Network Research, pp. 137-164. Academic Press, New York (1979)
27. Rogers, E.M.: Diffusion of Innovations [1962]. The Free Press, New York (1995)
28. Rogers, E.M., Ascroft, J.R., Röling, N.: Diffusion of Innovation in Brazil, Nigeria, and India. Unpublished Report. Michigan State University, East Lansing (1970)
29. Rogers, E.M., Kincaid, D.L.: Communication Networks: Towards a New Paradigm for Research. The Free Press, New York (1981)
30. Simonsen, I., Eriksen, K.A., Maslov, S., Sneppen, K.: Diffusion on complex networks: a way to probe their large-scale topological structures. Physica A **336** 163–173 (2004)
31. Simonsen, I.: Diffusion and networks: a powerful combination. Physica A **357** 317–330 (2005)
32. Skillnäs, N.: Modified innovation diffusion. A way to explain the diffusion of cholera in Linköping in 1866. A study of methods. Geografiska Annaler B: Human Geography **81** 243–260 (1999)
33. Skvoretz, J.: Salience, heterogeneity, and consolidation of parameters: civilizing Blaus primitive theory. American Sociological Review **48** 360–375 (1983).

34. Stauffer D., Solomon, S.: Physics and mathematics applications in social science. In: Mayers, R.A. (ed): Encyclopedia of Complexity and Systems Science. Springer, New York, pp. 6804-6810 (2009). Preprint: arXiv:0801.01121v1 [physics.soc-ph]
35. Valente, T.W.: Network Models of the Diffusion of Innovations. Hampton Press, Creskill (1995)
36. Valente, T.W.: Social network thresholds in the diffusion of innovations. Social Networks **18** 69–89 (1996)
37. Valente, T.W.: Network models and methods for studying the diffusion of innovations. In: Carrington, P.T., Scott, J. and Wasserman, S. (eds.) Models and Methods in Social Network Analysis, pp. 98-116. Cambridge University Press, New York (2005)
38. Verhulst, P.F.: Notice sur la loi que la population suit dans son accroissement. Corresp. math. phys. **4** 113–121 (1838)
39. Watts, D.J., Strogatz S.H.: Collective dynamics of small word networks. Nature **393** 440–442 (1998)

Credits

Appendix: Academic Network DHA Indices

Code	University	D_i	H_i	A_i	D'_i	H'_i	A'_i	Sign	ω_i
ROS	Røskilde	0.111	0.000	0.867	−0.048	−0.323	0.389	− − +	+0.010
MAG	Magdeburg	0.111	0.410	0.487	−0.048	0.087	0.009	− + +	+0.028
ERF	Erfurt	0.111	0.571	0.417	−0.048	0.248	−0.061	− + −	+0.080
PRA	Prague	0.111	0.410	0.396	−0.048	0.087	−0.082	− + −	−0.025
ESZ	Esztergom	0.111	0.000	0.908	−0.048	−0.323	0.430	− − +	+0.034
WIE	Vienna	0.111	0.000	0.287	−0.048	−0.323	−0.191	− − −	−0.324
RGB	Regensburg	0.125	0.727	0.367	−0.034	0.404	−0.111	− + −	+0.149
KOL	Cologne	0.125	0.700	0.604	−0.034	0.377	0.126	− + +	+0.271
STR	Strasbourg	0.143	0.571	0.158	−0.016	0.248	−0.320	− + −	−0.051
DUB	Dublin	0.125	0.000	0.900	−0.034	−0.323	0.422	− − +	+0.038
OXF	Oxford	0.125	0.000	0.333	−0.034	−0.323	−0.145	− − −	−0.290
CAM	Cambridge	0.125	0.000	0.458	−0.034	−0.323	−0.020	− − −	−0.218
PAR	Paris	0.143	0.812	0.104	−0.016	0.489	−0.374	− + −	+0.082
ANG	Angers	0.143	0.000	0.567	−0.016	−0.323	0.089	− − +	−0.145
DIJ	Dijon	0.167	0.571	0.225	0.008	0.248	−0.253	+ + −	+0.002
AVI	Avignon	0.167	0.000	0.246	0.008	−0.323	−0.232	+ − −	−0.316
MON	Montpellier	0.167	0.000	0.408	0.008	−0.323	−0.070	+ − −	−0.222
TOU	Toulouse	0.143	0.250	0.492	−0.016	−0.073	0.014	− − +	−0.043
LIS	Lisbon	0.111	0.000	0.712	−0.048	−0.323	0.234	− − +	−0.079
SAL	Salamanca	0.111	0.000	0.483	−0.048	−0.323	0.005	− − +	−0.211
SEV	Seville	0.111	0.000	0.479	−0.048	−0.323	0.001	− − +	−0.214
VAL	Valencia	0.125	0.571	0.458	−0.034	0.248	−0.020	− + −	+0.112
AST	Asti	0.200	0.750	0.658	0.041	0.427	0.180	+ + +	+0.374
MIL	Milan	0.200	0.000	0.546	0.041	−0.323	0.068	+ − +	−0.124
GEN	Genoa	0.200	0.000	0.283	0.041	−0.323	−0.195	+ − −	−0.275
PIS	Pisa	0.250	0.727	0.650	0.091	0.404	0.172	+ + +	+0.385
BOL	Bologna	0.250	0.400	0.271	0.091	0.077	−0.207	+ + −	−0.023
PAD	Padua	0.250	0.000	0.458	0.091	−0.323	−0.020	+ − −	−0.145
VEN	Venice	0.250	0.000	0.671	0.091	−0.323	0.193	+ − +	−0.023
RIM	Rimini	0.333	0.625	0.542	0.174	0.302	0.064	+ + +	+0.312
FIR	Florence	0.250	0.250	0.083	0.091	−0.073	−0.395	+ − −	−0.217
SIE	Siena	0.250	0.400	0.333	0.091	0.077	−0.145	+ + −	+0.013
PER	Perugia	0.200	0.400	0.317	0.041	0.077	−0.161	+ + −	−0.025
ASS	Assisi	0.167	0.625	0.271	0.008	0.302	−0.207	+ + −	+0.059
TOD	Todi	0.167	0.400	0.250	0.008	0.077	−0.228	+ + −	−0.083
ROM	Rome	0.167	0.823	0.958	0.008	0.500	0.480	+ + +	+0.570
ASC	Ascoli	0.143	0.400	0.416	−0.016	0.077	−0.062	− + −	−0.001
CHI	Chieti	0.143	0.812	0.250	−0.016	0.489	−0.228	− + −	+0.141
DJU	Djurazci	0.125	0.000	0.896	−0.034	−0.323	0.418	− − +	+0.035
CIV	Città	0.125	0.727	0.150	−0.034	0.404	−0.328	− + −	+0.024
BAR	Barletta	0.125	0.571	0.800	−0.034	0.248	0.322	− + +	+0.309
NAP	Napoli	0.125	0.700	0.683	−0.034	0.377	0.205	− + +	+0.317
REG	Reggio	0.125	0.000	0.604	−0.034	−0.323	0.126	− − +	−0.133
PAL	Palermo	0.143	0.000	0.583	−0.016	−0.323	0.105	− − +	−0.135

Interlocking Communication
Measuring Collaborative Intensity in Social Networks

Klaus Stein and Steffen Blaschke

Abstract Research on collaboration in social networks is largely restricted by a lack of longitudinal data. Approximations of collaborative intensity necessarily rely on the width of collaboration, such as the number of papers two scientists have coauthored. In contrast, we discuss precise measures of collaborative intensity with respect to not only the width but also the depth of collaboration. Based on empirical data of four social networks, we compare a widely-used approximation with our own measures of collaborative intensity. We find that the quality of the approximation varies with the type of social networks.

1 Introduction

The study of social networks is an almost natural approach to collaboration. A network comprises of a set of nodes and a set of ties representing some relationship between the nodes. The nodes in social networks are most commonly individuals, organizations, or societies [17, 4, 5, 20], and the ties often represent collaboration as the particular type of relationship between the nodes. This is true for coauthorship in the well-known Erdős collaboration graphs [10, 1] as well as collective decisions in the infamous Enron email corpus [7]. Other examples are easy to come by, just consider collaboration in terms of the scientific research at universities, the corporate development of consumer products, or the world-wide fight against injustice and poverty.

Klaus Stein
Laboratory for Semantic Information Technology, University of Bamberg, Germany,
e-mail: klaus.stein@uni-bamberg.de

Steffen Blaschke
Chair of Organization and Management, University of Hamburg, Germany,
e-mail: steffen.blaschke@wiso.uni-hamburg.de

The intensity of collaboration is among the standard measures in social network analysis, typically displayed by the weight of ties [21, 26]. Unfortunately, collaborative intensity is mostly a simple approximation such as the number of coauthored papers or sent emails. The simplicity of this approximation is not so much a shortcoming of research but rather a matter of data restriction. For example, data on the coauthorship of scientific papers allows for no other measure of collaborative intensity than the number of papers two or more scientists have coauthored. That is to say, the intensity of collaboration on a single paper does not reflect in the available data. In contrast, the amount of data available in collaborative authoring systems such as Wikipedia or social networking sites such as Facebook, MySpace, or Twitter allows for precise measures of collaborative intensity instead of simple approximations.

In the following, we remedy this shortcoming with a measure of collaborative intensity based on empirical data of four social networks we obtain from collaborative authoring systems in respective organizations. First, we present a widely-used algorithm to approximate collaborative intensity in social networks [16, 15]. We introduce the concept of interlocking communication [2] to determine collaboration in depth and width, and visually compare representations of social networks according to the various measures. Finally, we statistically revisit these measures and discuss the implications thereof.[1]

2 Research on Collaboration Networks

Beginning in the 1960s, collaboration in coauthorship networks receives ample research year after year (for original works and reviews thereof, cf. [23, 24, 14, 19]). The now established research field of scientometrics devotes particular attention to networks of scientists who have coauthored one or more papers. Yet, collaboration in coauthorship networks is certainly not restricted to scientists. It extends to all individuals who have coauthored any kind of document, be it scientific papers, project reports, programming code, to-do lists, or and emails. Scientific coauthorship networks nevetheless remain a primary source for research on collaboration, mainly because large-scale data is publically available.

Based on data of published scientific papers (e. g., MEDLINE and NCSTRL), Newman [16] constructs a weighted collaboration network by using the number of papers each pair of scientists has coauthored. In the absence of other available data, he estimates the intensity of collaboration by the reciprocal of the number of coauthors. This follows the idea that a scientist needs to split the time she puts into one paper between her coauthors so that, on average, the collaboration with one of them decreases with the

[1] This work was supported by the *Volkswagen Stiftung* through Grant No. II/82 509.

number of other coauthors. Newman admits that this is a coarse assumption due to the lack of more detailed data. Moreover, he assumes that scientists who coauthor many papers, on average, know each other better than those who collaborate on fewer papers. Following Newman [15], the collaborative intensity between two authors A and B on a set of papers P is defined as

$$w(A \leftrightarrow B) := \sum_{\substack{p \in P \\ A \lessdot p \\ B \lessdot p}} \frac{1}{n_p - 1}, \tag{1}$$

$$U \lessdot p :\Leftrightarrow U \text{ is author of paper } p,$$

$$n_p :\Leftrightarrow \text{ the number of authors of paper } p.$$

Weighted collaboration networks are commonplace in today's research (e. g., [11, 18]), but they barely extend the approximation that Newman proposes. The availability of longitudinal data then renders coarse assumptions obsolete, and we may finally replace any approximation by a true measure of collaborative intensity.

3 From Communication to Collaboration

Collaboration comes about communication. We may define communication as "reproducing at one point either exactly or approximately a message selected at another point" [22, p. 379], as "the transmission of a meaning, assuming that we are capable of understanding one another as concerns each of these words (transmission, meaning, etc.)" [6, p. 330], or as "a synthesis of three different selections, namely, selection of *information*, selection of *utterance* of this information, and a selective *understanding or misunderstanding* of this utterance and its information" [12, p. 252]. The most important aspect of these definitions in terms of social networks is that communication establishes a relationship between a sender and a receiver [22], an addressor and an addressee [6], or an alter and an ego [13]. Therefore, we address communication firstly as a tie between two nodes.

Let us illustrate communication by means of an example. Consider two scientists who coauthor a paper using a word processor (e. g., Microsoft Word or Open Office Writer). From the first draft to the final version, both authors consecutively revise the paper in response to each other such that each one of their revisions completes a single communication event. If we define collaboration as the joint effort of two or more entities in an intellecutal endeavor, then the history of communication events—one revision after another—is evidence of continuous communication or, in other words, collaboration between the two scientists.

The case in point is easy to follow for any pair of scientists collaborating on a single paper. Some papers are short and straight-forward pieces of writing with little need for consecutive revisions, others are long and elaborate contemplations with hundreds or thousands of revisions before publication. Therefore, collaboration on short papers tends to be less intense, and vice versa. This reasoning holds true for any pair of scientists collaborating on more than just a single paper, too. However, collaborative intensity now reflects in not just a single history of communication events but in as many as there are papers by any pair of scientists.

Unfortunately, the fact that many scientific papers spot more than two coauthors complicates matters. If we think of a communication event as a direct response of a scientist to the previous writing of another scientist, then we disregard all indirect responses of the scientist to the previous writings of others. And if we disregard the history of communication events that lead up to the paper as such, then we deny individuals and organizations any form of memory. To remedy this shortcoming, we rely on the concept of interlocking communication [2] which builds upon the above definition of communication as emerging from ego's selections of information, utterance, and alter's selective understanding thereof. For two scientists collaborating on a single paper, communication interlocks at exactly each communication event in reference to previous ones.

The following example illustrates interlocking communication in more detail. Consider two papers p_1 and p_2 with revision histories

$$H_1 = [a, b, a, b, a, b] \quad \text{and} \quad H_2 = [a, a, a, b, b, b],$$

where a is a revision of scientist A and b is a revision of scientist B. Obviously, the number of turns in communication are higher in H_1 than in H_2. Indeed, there is no way to tell whether or not A even noticed that B is working on p_2. In addition, consider another paper p_3 with the history

$$H_3 = [a, c, b, d, a, d, b, d, c, a, c, b, c, d],$$

where c and d are revisions of another two scientists C and D. Here, A and B do not directly respond to each other at all, although p_3 is certainly the result of collaboration among all coauthors. Communication between A and B takes place in reference to all previous revisions, despite the revisions of C and D in between.

In order to highlight the collaboration between A and B, we ignore all the revisions of others than these two scientists and get

$$H_3 = [a, b, a, b, a, b].$$

The directed tie from respective nodes A to B in a social network then yields a weight of two interlocking communications ($A \to B = 2$), whereas the directed tie from B to A yields a weight of three ($B \to A = 3$). For paper

p_2, we set a tie from B to A (at a weight of 1) but none from A to B. The weight of ties in terms of interlocking communications effectively addresses the issue of collaborative intensity. Still, we need to elaborate on the depth and width of collaboration.

Obviously, interlocking ignores the *amount of text* contributed by each author. This is legitimate as we want to measure the *amount of collaboration* based on the degree of interaction and not the *amount of contribution* of the single author to the resulting text.

4 Collaborative Intensity

Interlocking communication tells us about the extent of a joint effort between any two nodes on a given intellectual endeavor, that is, the the collaborative intensity of two coauthors A and B on a single paper. Since collaboration in scientific research usually spans several themes and topics (i. e., papers), we see interlocking communications between A and B in a number of revision histories. Therefore, we must consider

- the collaborative depth of $A \leftarrow B$: the number of interlocking communications from B to A on a single paper, and
- the collaborative width of $A \leftarrow B$: the number of papers with interlocking communications from B to A.

With $\mathrm{il}_p(A \leftarrow B)$ being the interlocking communication weight for B with respect to A on paper p, we define interlocking communication from B to A on a set of papers P:

$$\text{add:} \qquad \mathrm{il}^{+}(A \leftarrow B) := \sum_{p \in P} \mathrm{il}_p(A \leftarrow B) \tag{2}$$

$$\text{max:} \qquad \mathrm{il}^{\top}(A \leftarrow B) := \max_{p \in P} \{\, \mathrm{il}_p(A \leftarrow B) \,\} \tag{3}$$

$$\text{paper:} \qquad \mathrm{il}^{\#}(A \leftarrow B) := |\{\, p : \mathrm{il}_p(A \leftarrow B) > 0 \,\}| \tag{4}$$

$$\text{expr:} \qquad \mathrm{il}^{\lceil k}(A \leftarrow B) := \sqrt[k]{\sum_{p \in P} (\mathrm{il}_p(A \leftarrow B))^k} \tag{5}$$

il^{+} (add) provides a balance between collaborative depth and width in that it collects all interlocking communications across papers. $\mathrm{il}^{\top}$ (max) enforces the depth of collaboration. It simply considers the maximum of interlocking communications across papers but ignores the number of papers that exhibit interlocking to begin with. In contrast, $\mathrm{il}^{\#}$ (paper) enforces the width of collaboration. It is sensitive to the number of papers any pair of scientists coauthors but ignores the depth of collaboration on these papers. $\mathrm{il}^{\lceil k}$ (expr) is an intermediate measure that allows us to finetune the importance

of collaborative depth as oppossed to width by changing k. For

$$\begin{aligned} k > 1: &\quad \mathrm{il}^{\lceil k} \text{ favors depth,} \\ k = 1: &\quad \mathrm{il}^{\lceil 1} = \mathrm{il}^{+}, \\ k < 1: &\quad \mathrm{il}^{\lceil k} \text{ favors width.} \end{aligned}$$

Computing any one of these measures for each pair of coauthors allows us to build social networks of authors (nodes) connected by their collaborative intensity at the weight of their interlocking communication (ties). Moreover, the collaborative effort of a single author reflects in her weighted node degree, that is, the sum of the weights of her interlocking communications in the social network. It is furthermore possible (and certainly makes sense) to distinguish the collaborative depth and width of an author by addressing the intensity of collaboration with another author as well as the number of other authors she collaborates with. For now, and in addition to our above tie-based measures, we stick with the cummulative weighted degrees in order to compare these to the Newman approximation of collaborative intensity.

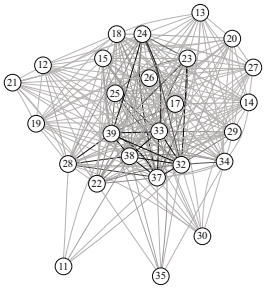
5 Case Studies

In order to illustrate collaborative intensity in social networks, we rely on data of four collaborative authoring systems. These systems are altogether corporate wikis which, similar to a word processor, keep track of each and every revision to a document. The uncontested success of the free online encyclopedia Wikipedia is evidence of the ease and elegance of collaboration with the help of wikis.

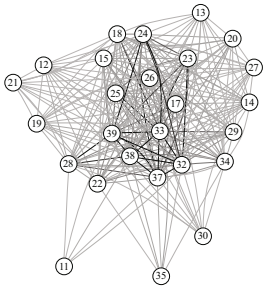
Four German organizations provide us unrestricted access to their corporate wikis. All of them employ a standard installation of MediaWiki, the same software that Wikipedia uses. Their technological environments are similar, whereas their organizational environments differ in management attitude towards the respective wiki. In-depth case studies of the organizations are provided by [25].

The first organization is the European market leader for one-stop solutions and services in mobile or proximity marketing (e. g., bluetooth hotspots). Its *startup wiki* supports all organizational functions, ranging from research and development to marketing and sales. From day one of the installation, the wiki is the primary collaboration system and, indeed, the only content management system to date.

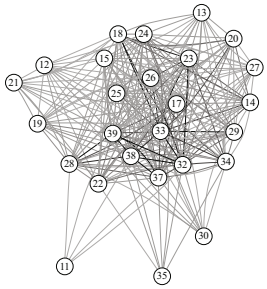
The second organization is a public center of excellence that lends support for information and communication technologies to small and medium enterprises. Its *facilitation wiki* features mostly research articles, project reports,



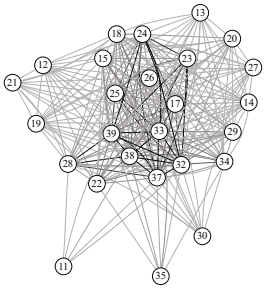
(a) sup: newman



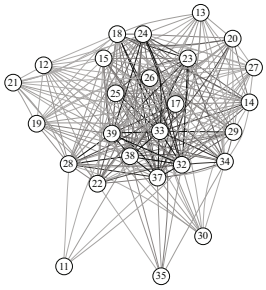
(b) sup: il^{+}



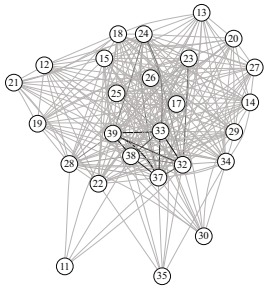
(c) sup: $il^{\top}$



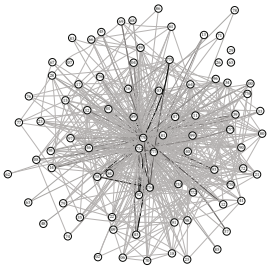
(d) sup: $il^{\#}$



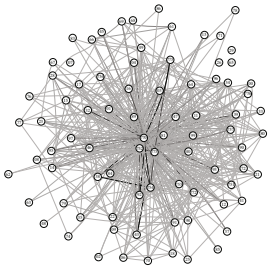
(e) sup: $il^{\lceil 2}$



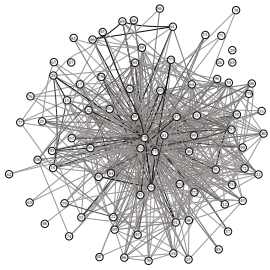
(f) sup: $il^{\lceil 0.5}$



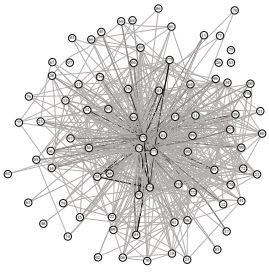
(g) fac: newman



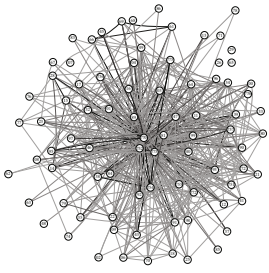
(h) fac: il^{+}



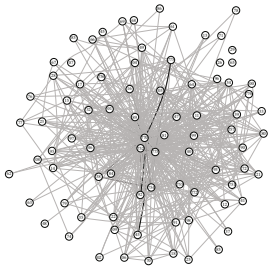
(i) fac: $il^{\top}$



(j) fac: $il^{\#}$

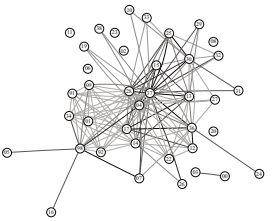


(k) fac: $il^{\lceil 2}$

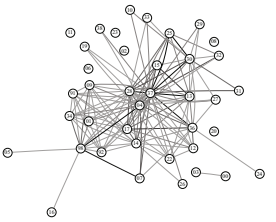


(l) fac: $il^{\lceil 0.5}$

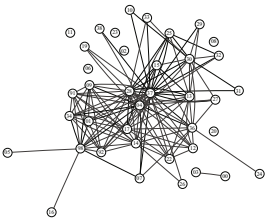
Fig. 1: Social Networks of the *Startup* (sup) and *Facilitation* (fac) Wikis



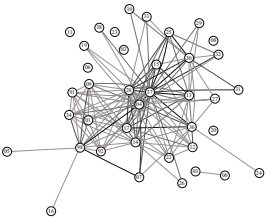
(a) new: newman



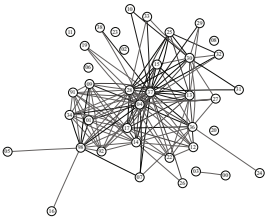
(b) new: il^{+}



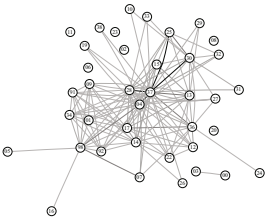
(c) new: $il^{\top}$



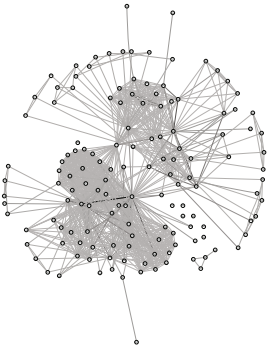
(d) new: $il^{\#}$



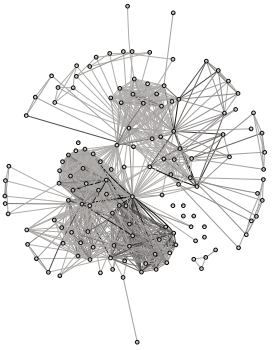
(e) new: $il^{\lceil 2}$



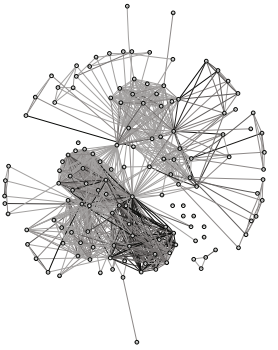
(f) new: $il^{\lceil 0.5}$



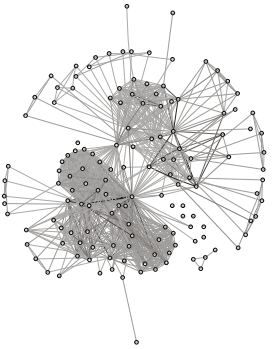
(g) uni: newman



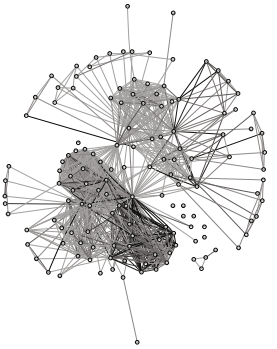
(h) uni: il^{+}



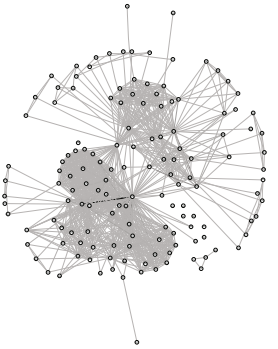
(i) uni: $il^{\top}$



(j) uni: $il^{\#}$



(k) uni: $il^{\lceil 2}$



(l) uni: $il^{\lceil 0.5}$

Fig. 2: Social Networks of the *News* (new), and *Students* (uni) Wikis

and later publications thereof. The wiki is maintained as a dedicated project, and employees are enforced to generate a certain amount of input per anno.

The third organization is a public news agency with a focus on online broadcasts. For the most part, its *news wiki* comprises production manuals, frequently asked questions, and official meeting protocols, complementing an editorial content management system. Contrary to the other wikis, this wiki allows anonymous edits, that is to say, we are not always able to distinguish single users but have to pool IP-ranges to prototypic user-groups based on their function (e. g., journalists, illustrators, interns).

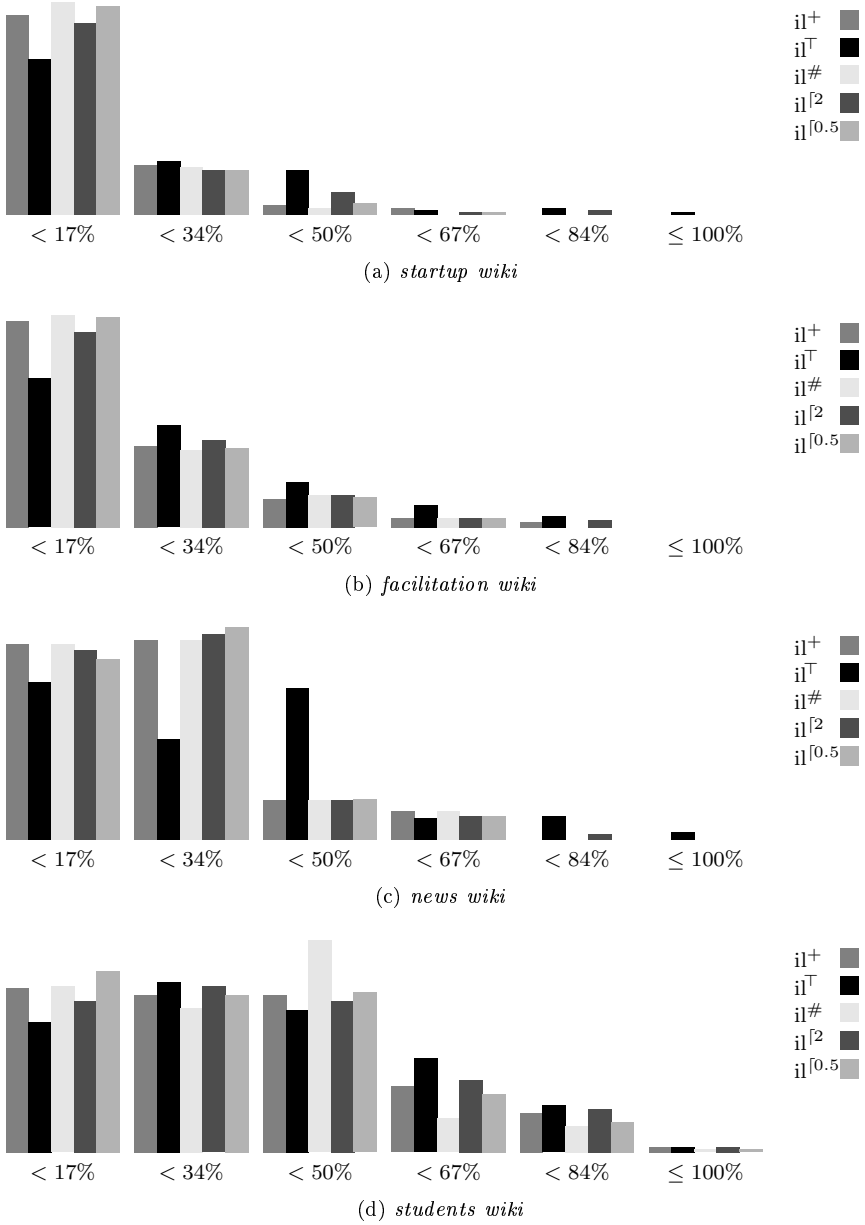
The fourth and last organization is a loosely-coupled community of students within a faculty of media science. Its *students wiki* centers around the self-assigned project to build a collection of encyclopedic articles on topics such as e-learning and knowledge management. The students use the wiki as a collaboration platform for dedicated courses. Therefore, the body of encyclopedic articles grows with each semester of new authors arriving and old ones leaving.

Figures 1 and 2 provide six different measures of collaborative intensity for each one of the four social networks on a separate row. The nodes represent coauthors and the ties their collaboration on at least one page in the wiki. The darker a tie, the more intense is the collaboration between coauthors. The basic layout is the Newman collaboration network as introduced above. For easy comparison, we maintain the position of nodes in the basic layout and only change the weight of the ties according to the results of our different computations of collaborative intensity.

The different measures of collaborative intensity highlight one or the other particularity about each one of the social networks. For example, see first the Newman approximation in Figure 1g. Collaboration appears intense (i. e., with dark ties) only for a few pairs of coauthors at the center of the network. Now, compare this finding to $\mathrm{il}^{\top}$ in Figure 1i. Collaborative intensity is at higher levels (i. e., ties are overall darker) since the measure favors the depth of collaboration. In particular, there is a triad of nodes, located in the middle of the upper periphery of the network, that shows only if we take the history of communication events into account. The collaboration among these three nodes is barely noticable in the Newman approximation since it concerns mostly a single page. However, this page comes about very intense collaboration. The triad also shows in $\mathrm{il}^{\lceil k}$ at $k = 2$, which also favors the depth of collaboration.

6 Evaluating Collaborative Intensity

How does the Newman approximation of collaborative intensity compare to our computations based on the concept of interlocking communication? Obviously, all social networks comprise of the same nodes and ties but differ in the



(a) *startup wiki*

(b) *facilitation wiki*

(c) *news wiki*

(d) *students wiki*

Fig. 3: Tie Weight Rank Offsets

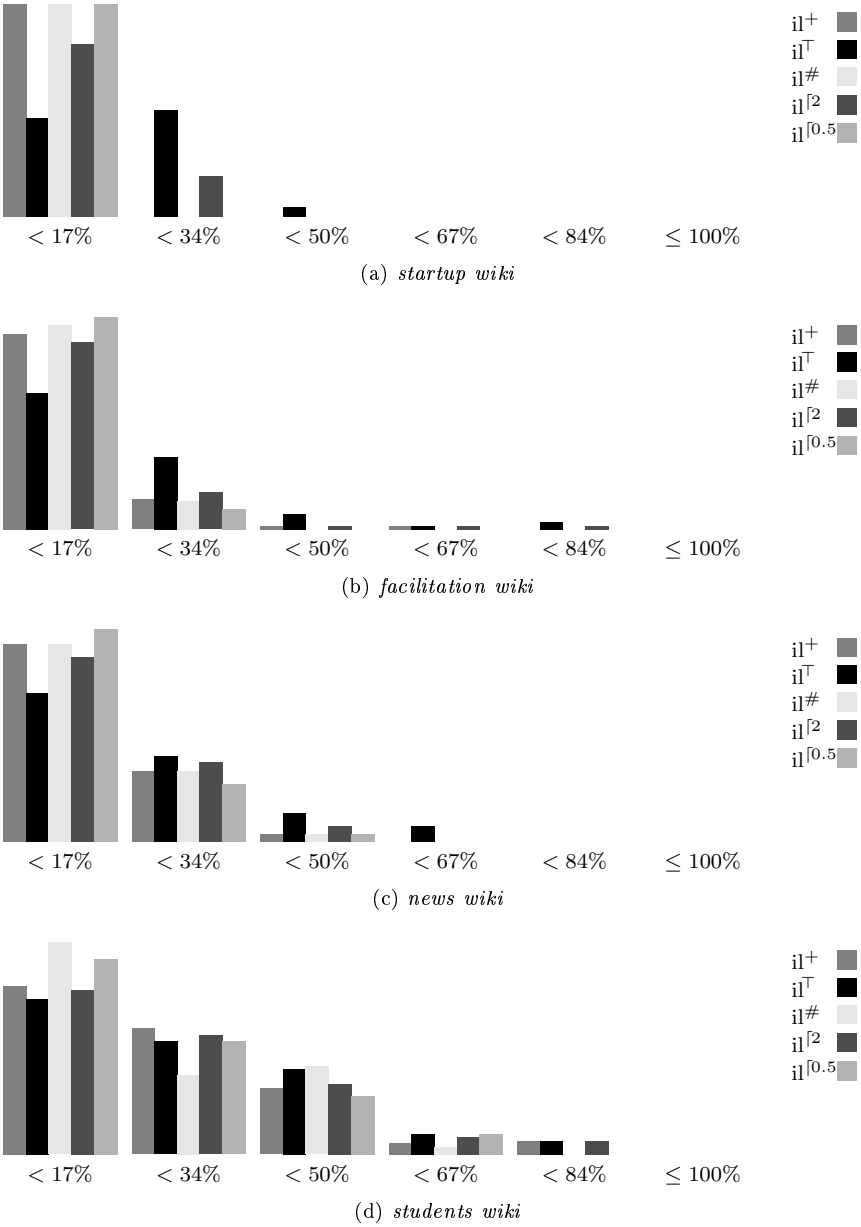


Fig. 4: Node Degree Rank Offsets

Table 1: Average Tie Weight Rank Offsets

	il^{+}	$il^{\top}$	$il^{\#}$	$il^{\lceil 2}$	$il^{\lceil 0.5}$
startup wiki	12.2 %	20.1 %	10.8 %	14.9 %	10.6 %
facilitation wiki	15.6 %	22.7 %	14.7 %	17.2 %	14.4 %
news wiki	21.1 %	29.8 %	21.0 %	21.7 %	20.5 %
students wiki	31.4 %	34.6 %	30.2 %	32.6 %	29.8 %

Table 2: Average Node Degree Rank Offsets

	il^{+}	$il^{\top}$	$il^{\#}$	$il^{\lceil 2}$	$il^{\lceil 0.5}$
startup wiki	4.4 %	14.2 %	4.4 %	9.2 %	3.0 %
facilitation wiki	8.4 %	13.8 %	7.7 %	11.3 %	5.4 %
news wiki	8.6 %	10.9 %	8.4 %	10.1 %	7.1 %
students wiki	14.2 %	16.3 %	13.5 %	15.4 %	12.4 %

weight of ties. As a direct comparison of these weights is pointless, we compare the rankings of ties, that is, we sort all ties of one network by weight and compare their positional change with respect to the Newman approximation. This provides us with a network similarity measure.

Figure 3c, for example, shows that with the $il^{\top}$ measure of collaborative intensity in place (black bars) roughly 35 % of the ties change less than 17 % but more than 40 % change more than 34 % in rank. The same holds true for the other networks: the ties in the $il^{\top}$ network change the most such that $il^{\top}$ takes the opposite to the Newman approximation in that it favors the depth of collaboration. This is unsurprising since the approximation is based on the number of coauthored papers which favors the width of collaboration.

The average change in tie ranks for the different measures supports this finding as shown in Table 1. For all networks, $il^{\top}$ brings about the biggest change, followed by $il^{\lceil 2}$ which also favors the depth of collaboration. Again, similar changes are also visible in respective author degree rankings (see Figure 4 and Table 2), where both $il^{\top}$ and $il^{\lceil 2}$ show the largest differences to the Newman collaboration network.

Let us point out some more particularities. Notice the distribution of rank offset in Figure 3c as it compares to Figure 3d. While the rank offset of ties in the *startup wiki* remains somewhat close to the Newman approximation, the *students wiki* yields a completely different picture with the rank offset drastically removed. Recall that the *students wiki* reflects a social network of coauthors who collaborate intensely on but a few documents. In such a case, the Newman approximation of collaborative intensity is simply not a good measure, whereas any one of our measures accurately captures both the width and depth of collaboration.

7 Filtering

Interlocking networks as shown in Figures 1, 2 connect two authors whenever they have at least one edit on at least one common paper. Therefore, interlocking networks are structurally equivalent to coauthorship networks, except for the fact that they have directed ties.

On the notion that collaboration requires more than a single or even a few interactions, we may demand that two authors interact more than $w^{\top}$ times on a single paper or that they have interacted on more than $w^{\#}$ common papers, otherwise they are not connected by a tie. In the simplest of all applications, we then remove all ties below or equal to a weight $w^{\top}$ from the interlocking graph $\mathrm{il}^{\#}$ or a weight $w^{\#}$ from the interlocking graph $\mathrm{il}^{\top}$, respectively. In other words, we first construct an interlocking graph (il^{+}, $\mathrm{il}^{[2}$, $\mathrm{il}^{\top}$, ...) only to filter all ties below a given weight, thereby cutting weak ties and keeping strong ones.

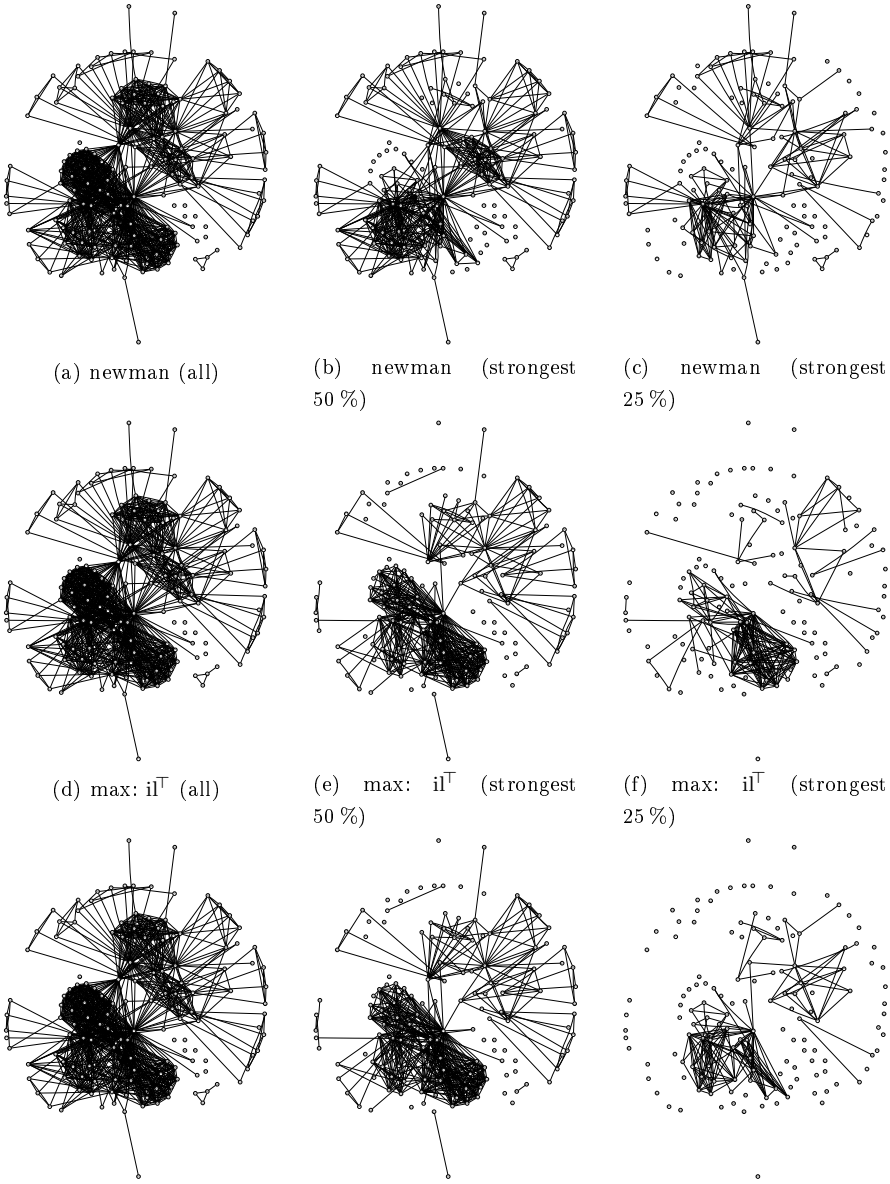
The top row of Figure 5 shows the Newman graph of the *students wiki* in full (5a), with 50 % (5b), and 75 % (5c) of the weak ties removed.[2] A quick comparison of these three graphs to the below max ($\mathrm{il}^{\top}$) and paper ($\mathrm{il}^{\#}$) graphs in Figure 5d to 5i points out the differences between the Newman approximation and our interlocking measures we discuss in the previous section.

Both the max ($\mathrm{il}^{\top}$) graph and paper ($\mathrm{il}^{\#}$) graph fall apart in (almost distinct) components as we increase the filter criteria. The reason is that the students who are using this wiki change with each semester. Most commonly, they join a course, work on a subset of pages over the course of the semester, and then leave. Different course exhibit different sets of students with only little overlap, which effectively decreases the chance of interlocking communication among students. Looking at collaborative intensity either way reveals these social dynamics, whereas commonplace measures such as the Newman approximation hide it.

Removing ties below or equal to a given weight allows for the comparison of different networks, which reveals other social dynamics of interest. Figures 6 and 7 show the *students wiki* in contrast to the *facilitation wiki*.

The *facilitation wiki* highlights a single author at the center of the network who interacts with many others across many pages, which the filter with respect to the page graph ($\mathrm{il}^{\#}$) only underlines. The depth of collaboration ($\mathrm{il}^{\top}$), however, is comparibly low with different social dynamics in place, though the central author remains the same. This author is in fact the project leader responsible for the wiki. Our interviews confirm that she gets calls for help on many occasions (i. e., concerning many pages) from many other members of the organization. Moreover, the primary use of the wiki is more

[2] More precisely, we sort ties according to their weight and keep only those strong ties with a weight above the 50 % or 25 % tie weight. Since some ties frequently have the same weight, we may remove slightly more than 50 % or 75 % of the actual ties. Still, we prefer this procedure to randomly choosing which ties to remove from the set of equal-weight ties.



(a) newman (all)

(b) newman (strongest 50 %)

(c) newman (strongest 25 %)

(d) max: $il^{\top}$ (all)

(e) max: $il^{\top}$ (strongest 50 %)

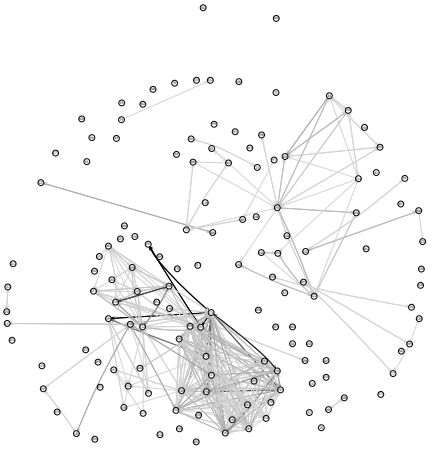
(f) max: $il^{\top}$ (strongest 25 %)

(g) page: $il^{\#}$ (all)

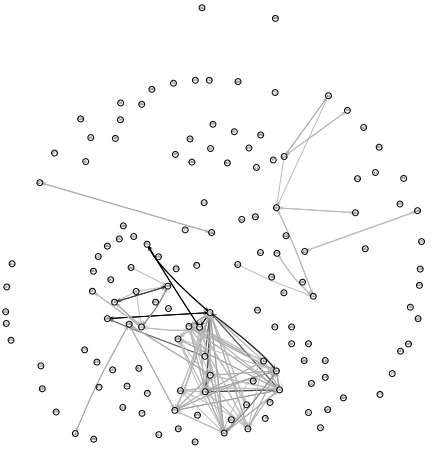
(h) page: $il^{\#}$ (strongest 50 %)

(i) page: $il^{\#}$ (strongest 25 %)

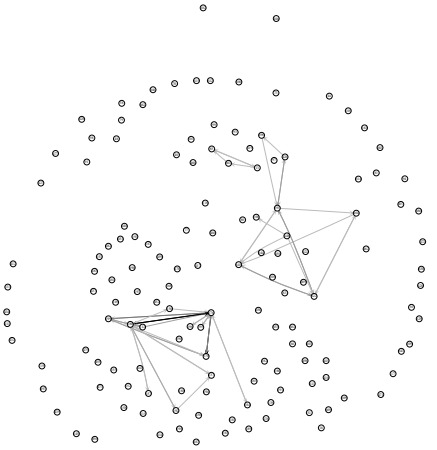
Fig. 5: *students wiki*, percentual filtered graphs.



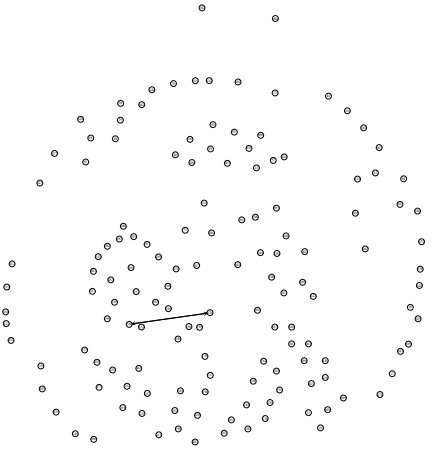
(a) max: $\mathrm{il}^{\top}$ ($w > 1$)



(b) max: $\mathrm{il}^{\top}$ ($w > 3$)

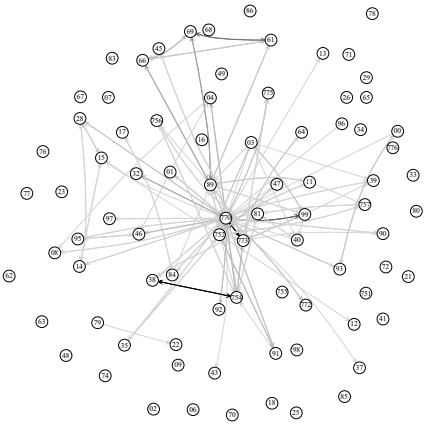


(c) page: $\mathrm{il}^{\#}$ ($w > 3$)

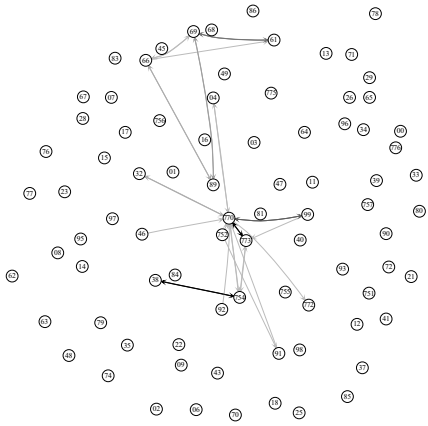


(d) page: $\mathrm{il}^{\#}$ ($w > 14$)

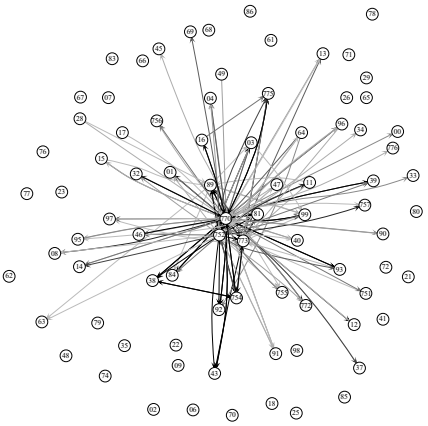
Fig. 6: *students wiki*, filtered graphs by fixed weights



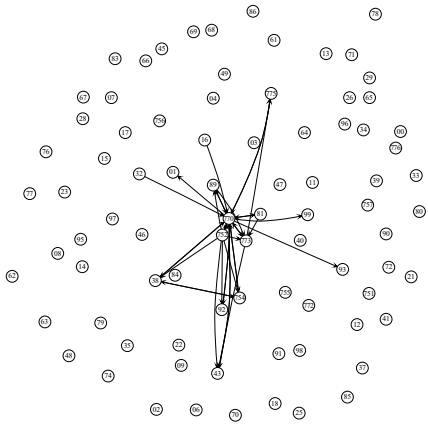
(a) max: $\mathrm{il}^{\top}$ $(w > 1)$



(b) max: $\mathrm{il}^{\top}$ $(w > 3)$



(c) page: $\mathrm{il}^{\#}$ $(w > 3)$



(d) page: $\mathrm{il}^{\#}$ $(w > 14)$

Fig. 7: *facilitation wiki*, filtered graphs by fixed weights

of an archive for already finished documents, not a tool to collaboratively develop ideas into full-blown pages.

For the *students wiki* we already know that the students work in groups on but a few pages that correspond to a respective course in the semester. They obviously work more *in* the wiki as their interlocking depth shows (compare Figures 6b to 7b).

Besides the quantitative data analysis, these two examples depict the importance of additional qualitative research without which appropriate interpretations are hard to come by, if not impossible at all. Social network analysis points out various phenomena, but only our background knowledge lets us tell the whole story.

Visual data mining helps us to detect such patterns, but it is certainly not the main aspect of filtering, mainly because filtering is highly dependent on the graph layout algorithm and thus may easily lead to misinterpretation. Here, a computer-assisted dynamic setup is of much aid. The researcher interactively alters the filter weight (e. g., using a slider) and the system reveals at which point the network falls into components or, the other way around, the system sequentially removes ties from weak to strong and thus provides snapshots at important transitions.

Filtering is a convenient way to identify the emergence of new components or highlight bridges that connect components, such as we see in Figure 5e where a triade of authors connect two components of the network. The weight at which these two components separate tells us just how intense members of each components (e. g., the authors of projects A and B) collaborate at minimum. Of course, filtered networks are open to standard measures from social network analysis. Rather than applying, for example, betweenness or closeness to filtered networks, we opt for weight-based modifications of these centrality measures.

8 Centrality in Weighted Networks

Among the standard measures in social network analysis is prestige or, in other words, the degree of a node, which is simply the sum of its ties to others in the network. We show in Section 4 how to use tie weights to compute weighted degrees in several ways, creating a set of interlocking-specific weight-aware measures by adapting the standard prestige measure to our needs. We are not the first to do so, not the least because weighted networks are a rather obvious concept, though they are hardly everyday business in social network analysis.

In the remainder of this section, we use the betweenness centrality measure to show how to apply weight-aware measures. We focus on the *startup wiki* since it is small enough in size to call upon nodes by their identification

number in the respective graphs. We compare the unweighted betweenness centrality [8] to length-scaled betweeness [3] and flow betweenness [9].

Betweenness measures the number of times a node A is on the shortest path from node B to node C, that is to say, the number of times A is *between* B and C. The path length is defined as the number of ties between B and C. Betweenness allows for a node to be interpreted as a gatekeeper or a bottleneck in the network, for instance. Length-scaled betweenness follows the idea that long paths should count less for the betweenness of a node than short ones, thus emphasing short distances.

Instead of counting the number of ties within a path as path length, we may as well annotate distances for each tie, thus computing the path length as sum of the distances of the participarting ties. Unfortunately, we cannot use the interlocking tie weights as distances. Larger weights should yield smaller distances just as stronger ties bring nodes closer together. Weight and distance are inversely related, so a hands-on approach is to compute the distance d_i of tie i from its interlocking link weight w_i as

$$d_i = w_{\max} + 1 - w_i, \qquad \text{with} \quad w_{\max} = \max_k \{w_k\}. \tag{6}$$

Freeman [9] addresses the same problem in a different manner. He, too, starts with the idea of betweenness but interprets ties as channels of communication and the tie weight as the capacity of this channel. His flow betweenness then takes a node A to be *between* two others B and C by the degree to which the maximum flow from B to C depends on A.

Table 3: *startup wiki*: betweenness, weight-aware length-scaled betweenness and weight-aware flow betweenness. All values rescaled to sum up to 1.

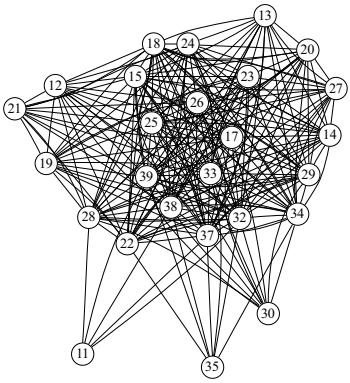
(a) betweenness for max: $\mathrm{il}^{\top}$

betweenness		l-sc. w. betw.		flow betw.	
$u39$:	0.1363	$u33$:	0.6117	$u33$:	0.1343
$u38$:	0.1363	$u39$:	0.1585	$u39$:	0.0943
$u32$:	0.1162	$u32$:	0.1465	$u32$:	0.0829
$u37$:	0.1162	$u37$:	0.0326	$u38$:	0.0718
$u33$:	0.0815	$u38$:	0.0232	$u37$:	0.0634
$u28$:	0.0593	$u28$:	0.0171	$u23$:	0.0514
$u26$:	0.0480	$u24$:	0.0026	$u28$:	0.0470
$u25$:	0.0480	$u26$:	0.0026	$u26$:	0.0434
$u34$:	0.0451	$u18$:	0.0026	$u18$:	0.0412
$u22$:	0.0391	$u34$:	0.0026	$u17$:	0.0400

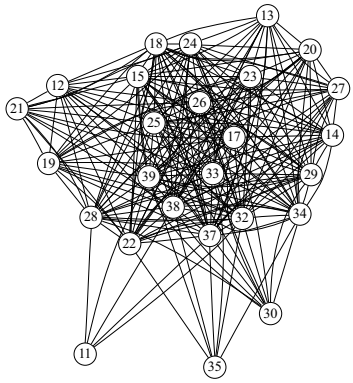
(b) betweenness for page: $\mathrm{il}^{\#}$

betweenness		l-sc. w. betw.		flow betw.	
$u39$:	0.1363	$u39$:	0.4227	$u39$:	0.1315
$u38$:	0.1363	$u33$:	0.1970	$u33$:	0.1276
$u32$:	0.1162	$u38$:	0.1400	$u32$:	0.1142
$u37$:	0.1162	$u32$:	0.1149	$u37$:	0.0960
$u33$:	0.0815	$u37$:	0.0790	$u38$:	0.0949
$u28$:	0.0593	$u24$:	0.0410	$u24$:	0.0701
$u26$:	0.0480	$u26$:	0.0034	$u25$:	0.0450
$u25$:	0.0480	$u25$:	0.0010	$u23$:	0.0446
$u34$:	0.0451	$u18$:	0.0010	$u17$:	0.0348
$u22$:	0.0391	$u30$:	0.0000	$u28$:	0.0335

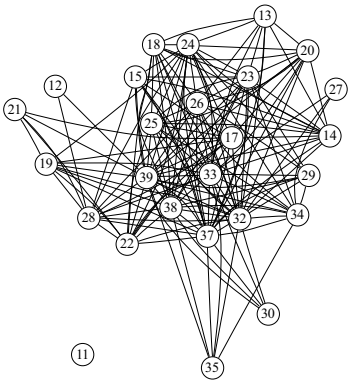
Tables 3a and 3b show the top ten authors according to their acscending betweenness scores with the unweighted betweenness in the left column, the length-scaled distance based betweenness (computed as introduced above)



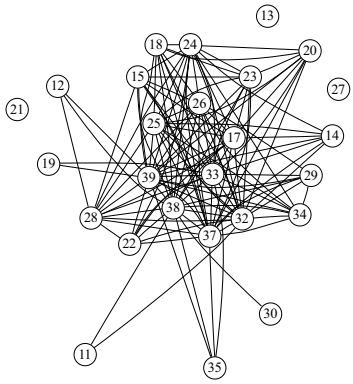
(a) max: $\mathrm{il}^{\top}$ (all)



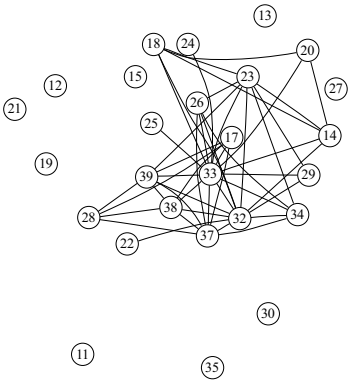
(d) page: $\mathrm{il}^{\#}$ (all)



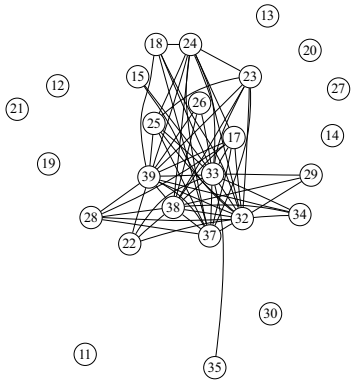
(b) max: $\mathrm{il}^{\top}$ (strongest 50 %)



(e) page: $\mathrm{il}^{\#}$ (strongest 50 %)



(c) max: $\mathrm{il}^{\top}$ (strongest 25 %)



(f) page: $\mathrm{il}^{\#}$ (strongest 25 %)

Fig. 8: *startup wiki*, percentual filtered graphs.

in the middle column and the flow betweenness in the right column. There are several observations worth making. First, note that author 33 is only in the fifth place with respect to his unweighted betweenness score, whereas he takes the top spot in the max graph ($il^{\top}$) and still the second spot in the page graph ($il^{\#}$). Both length-scaled betweenness and flow betweeness identify this node as more of a gatekeeper or bottleneck in collaboration than, for instance, author 39. Second, if we compare the actual scores of these two authors, we notice that lenght-scaled betweenness emphasizes the in-depth collaboration of author 33 (at a score close to four times the first runner-up, 0.6117 compared to 0.1585) as opposed to the more widely collaborative author 39 (at a score more than twice the first runner-up, 0.4227 compared to 0.197).

We also see that both weighted betweenness measures give similar results. While Freeman's flow betweenness is well founded on the capacity flow model, our decision to compute distances from weights by simple inversion is rather hands-on. Nevertheless, we show both measures since computing the flow betweenness is by far slower on large networks. For instance, for the facilitation network (85 users, 865 ties) it is factor 3700 (0.04 s to 148 s), for the students network (142 users, 1453 ties) factor 18000 (0.1 s to 30 min), all values measured on the R/sna[3] implementation. Additionally other measures like closeness or stress centrality can be made weight-aware using the same approach we use for the length-scaled distance-based betweenness.

Collaborative intensity may either take on depth or width, and it is most important to construct weighted networks and use weighted measures accordingly. If, on the one hand, the research question is to identify authors who collaborate in-depth on but a few pages, then weighted betweenness on a max graph certainly provides an reliable answer. On the other hand, if the research question is to identify authors who collaborate widely across many pages, then weighted betweenness on a page graph is the way to go.

9 Conclusion

Our computations of collaborative intensity take detailed data on the depth and the width of collaboration into account, something which Newman necessarily approximates. In the light of missing data on collaborative intensity, his approximations are of good quality for social networks with communication across many themes and topics. Networks with communication within but a few themes and topics, however, call for different measures such as ours.

We provide a first approach to collaborative intensity based on the concept of interlocking communication. The question which particular measure to choose depends on the needs of the research. In general, we tend to pick il^{+}

[3] http://www.r-project.org, http://cran.r-project.org/web/packages/sna

as a first choice since it balances the depth and the width of collaboration without the need of additional parameters. The next choice, of course, is to finetune either collaborative depth or width with the paramter k. Thus, we are able to determine the type of social network we are dealing with, for example, a tightly-integrated corporate network or a loosely-coupled research network. Furthermore, we show how to use weight filtering as well as weight-aware centrality measures for further analysis to reveal social dynamics in depth ($il^{\top}$) and width ($il^{\#}$).

All analysis and graphics in this paper are done with our own Wiki Explorator library (using Graphviz, R, and other open source applications). It is available for download as open source[4], but we also provide an online wiki analysis service based on this library[5].

References

1. Batagelj, V., Mrvar, A.: Some Analyses of Erdos Collaboration Graph. Social Networks **22**(2), 173–186 (2000)
2. Blaschke, S., Stein, K.: Methods and Measures for the Analysis of Corporate Wikis: A Case Study. In: Proceedings of the 58th Annual Conference of the International Communication Association. Montréal (2008)
3. Borgatti, S.P., Everett, M.G.: A Graph-Theoretic Perspective on Centrality. Social Networks **28**(4), 466–484 (2006)
4. Borgatti, S.P., Foster, P.C.: The Network Paradigm in Organizational Research: A Review and Typology. Journal of Management **29**(6), 991–1013 (2003)
5. Brass, D.J., Galaskiewicz, J., Greve, H.R., Tsai, W.: Taking Stock of Networks and Organizations: A Multilevel Perspective. Academy of Management Journal **47**(6), 795–817 (2004)
6. Derrida, J.: Margins of Philosophy. University of Chicago Press, Chicago, IL (1984)
7. Diesner, J., Frantz, T.L., Carley, K.M.: Communication Networks from the Enron Email Corpus. Computational & Mathematical Organization Theory **11**(3), 201–228 (2005)
8. Freeman, L.C.: Centrality in Social Networks: Conceptual Clarification. Social Networks **1**(3), 215–239 (1979)
9. Freeman, L.C., Borgatti, S.P., White, D.R.: Centrality in Valued Graphs: A Measure of Betweenness Based on Network Flow. Social Networks **13**(2), 141–154 (1991)
10. Grossman, J.W., Ion, P.D.F.: On a Portion of the Well-Known Collaboration Graph. Congressus Numerantium **108**, 129–131 (1995)
11. Liu, X., Bollen, J., Nelson, M.L., Van de Sompel, H.: Co-Authorship Networks in the Digital Library Research Community. Information Processing & Management **41**(6), 1462–1480 (2005)
12. Luhmann, N.: What is Communication? Communication Theory **2**(3), 251–259 (1992)
13. Luhmann, N.: Social Systems. Stanford University Press, Stanford, CA (1995)
14. Melin, G., Persson, O.: Studying Research Collaboration Using Co-Authorships. Scientometrics **36**(3), 363–377 (1996)
15. Newman, M.E.J.: Scientific Collaboration Networks. II. Shortest Paths, Weighted Networks, and Centrality. Physical Review E **64**(1), 016,132(1–7) (2001)

[4] http://wiki-explorator.rubyforge.org

[5] http://www.kinf.wiai.uni-bamberg.de/mwstat

16. Newman, M.E.J.: The Structure of Scientific Collaboration Networks. In: Proceedings of the National Academy of Sciences, pp. 404–409 (2001)
17. Oliver, A.L., Ebers, M.: Networking Network Studies: An Analysis of Conceptual Configurations in the Study of Inter-organizational Relationships. Organization Studies **19**(4), 549–583 (1998)
18. Opsahl, T., Panzarasa, P.: Clustering in Weighted Networks. Social Networks **31**(2), 155–163 (2009)
19. Osareh, F.: Bibliometrics, Citation Analysis and Co-Citation Analysis: A Review of Literature I. Libri **46**(3), 149–158 (1996)
20. Provan, K.G., Fish, A., Sydow, J.: Interorganizational Networks at the Network Level: A Review of the Empirical Literature on Whole Networks. Journal of Management **33**(3), 479–516 (2007)
21. Scott, J.: Social Network Analysis: A Handbook. Sage, London (1991)
22. Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana, IL (1949)
23. de Solla Price, D.J.: Networks of Scientific Papers. Science **149**(3683), 510–515 (1965)
24. de Solla Price, D.J.: A General Theory of Bibliometric and Other Cumulative Advantage Processes. Journal of the American Society for Information Science **27**(5–6), 292–306 (1976)
25. Stein, K., Blaschke, S.: Corporate Wikis: Comparative Analysis of Structures and Dynamics. In: K. Hinkelmann, H. Wache (eds.) Proceedings of the 5th Conference on Professional Knowledge Management, Lecture Notes in Informatics, pp. 77–86. Gesellschaft für Informatik, Bonn (2009)
26. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge, MA (1999)

The Structural Underpinnings of Policy Learning: A Classroom Policy Simulation

Stephen Bird

Abstract This paper investigates the relationship between the centrality of individual actors in a social network structure and their policy learning performance. In a dynamic comparable to real-world policy networks, results from a classroom simulation demonstrate a strong relationship between centrality in social learning networks and grade performance. Previous research indicates that social network centrality should have a positive effect on learning in other contexts and this link is tested in a policy learning context. Second, the distinction between collaborative learning versus information diffusion processes in policy learning is examined. Third, frequency of interaction is analyzed to determine whether consistent, frequent ties have a greater impact on the learning process. Finally, the data are analyzed to determine if the benefits of centrality have limitations or thresholds when benefits no longer accrue. These results demonstrate the importance of network structure, and support a collaborative conceptualization of the policy learning process.

In 1995, Texas officials began to consider competitive restructuring of the state's electricity industry. This effort included public utility commissioners, state legislators, private utility managers, environmentalists, consumer advocates, electricity cooperative representatives, and a variety of other stakeholders. While political concerns were certainly an important aspect of this

Stephen Bird,
Clarkson University,
PO Box 5750 - Humanities and Social Sciences
Potsdam, NY 13699, USA
e-mail: sbird@clarkson.edu

process, a critical component was that of policy learning.[1] All the various stakeholders needed to get "up to speed" on the various questions, promises, issues, and concerns that competitive restructuring held for their particular constituencies.[2] The learning process included trips by commissioners and legislators to the mid-Atlantic, Great Britain, and Canada to investigate different forms of restructuring. Stakeholders extensively communicated with influential voices within the Texas policy network, as well as national policy elites.[3]

Were the actors most central in this policy network the most likely to learn? What kind of communication patterns were most likely to improve their policy learning? What are the underlying mechanisms of the policy learning process? In this paper I argue that central structural positions of policy actors are an important component of policy learning, and that this learning is an interactive, collaborative process that goes beyond simple information transmission. This question is investigated by classroom simulation evidence meant to support state case studies analyzed in other writing. Policy learning is the deliberative process of gaining knowledge in the specialized areas of a policy realm by the stakeholders in a policy network. As experienced in Texas' electricity policy network, and in a wide variety of other policy networks, it is arguably a critical activity within any policy realm [23, 36, 45, 48, 54]. Adam and Kriesi emphasize that the core function of a policy network involves information exchange.

> The image of the policy network represents an intuitively comprehensible metaphor: *regular communication and frequent exchange of information* lead to the establishment of stable relationships between actors and to the coordination of their mutual interests. ([2007], 129, emphasis added)

Policy scholars know little about the role that social structure (i.e. communication and interaction patterns, often measured as "centrality") plays in a policy network context and the learning process of its stakeholders. Social networks have been found to be related to a variety of effects concerning performance, power, and influence [4, 20]. Various measurements of centrality have significant effects. Centrality has many variants, the simplest being

[1] "Policy learning" in this context does not mean policy adoption. This term assumes that proponents, opponents, and those without a policy opinion all must determine how policy - in this case electricity restructuring - will function within their personal and organizational beliefs and goals

[2] An extensive examination of this case study can be found in the dissertation of the author.

[3] "Policy networks" refer to the specialized and usually highly knowledgeable networks of stakeholders in any given policy domain such as highway policy, or health care reform, or electricity policy. This can include government, non-profit, academic, and various private sector actors within any policy domain. The network can be involved in the development, delivery, and also sometimes the process design leading to policy outputs

"degree centrality," measured simply by the amount of alters one interacts with.[4]

So why study social interactions within policy networks? One might argue that social network dynamics as studied by sociologists and organizational theorists should be generalizeable to policy networks. Policy networks, however, have unique aspects and characteristics that may make social network processes behave distinctively. Three important facets, widely discussed in the literature, are worth considering [26, 27, 37]. First, policy networks are highly bounded. Their participants have extensive technical knowledge of a scientific and/or regulatory nature in their policy arena that makes participation by others difficult. Second, network participants identify each other in large part through their formal roles within the network. This is, usually, the primary way in which identification occurs. Rather than identifying via location/geography, gender, race, or political identity, they do so by determining someone's role and/or organizational interest group: regulator, private sector (or some portion thereof), legislator, journalist, academic, consumer advocate, etc. Finally, policy actors are primarily concerned with policy outcomes. In generalized social situations, there may be a variety of different outcomes which actors may focus on. These can include wide-ranging economic rewards or resources, experiential activity, sexual interaction, or outcomes focused solely on the pleasure of social relations. While all of these kinds of social interactions can (and certainly do) occur within policy networks, the actors are predominantly concerned with regulatory or legislative results and outcomes.

I summarize these characteristics as follows:

1. Participation is limited to actors with extensive technical knowledge and background in the policy arena.
2. Interaction between actors focuses first or primarily on roles within network rather than other social or demographic identifiers.
3. Focus within policy networks is primarily set upon policy outcomes rather than other social outcomes.

Much research on policy networks has focused on policy learning processes from an institutional perspective ([13], 177). However, Sabatier ([47], 268) has explicitly called for policy network theorists to develop a more "coherent model of the individual." Zuckerman [55] has argued that a network perspective is useful for examining these mechanisms and providing the micro-level analysis for approaching these processes. Further, Hall [23] specifically applies the term "social learning" to the policy process to describe the social interaction processes that underlie policy learning (see also [22, 41, 42]). Part of the policy process is about power and influence, but another important

[4] Network analysis distinguishes between the actor being analyzed as "ego" and the persons they have contact with as "ties" or "alters."

arena is about learning, ideas, and discourse. Hall, using Heclo's terms, argues that social learning in policy networks is an even mix of "powering" and "puzzling."[5]

Individual level analysis in policy networks has not been done extensively because gaining reliable data from political elites is fraught with difficulty. One way to address this concern is to conduct policy network simulations to test individual level mechanisms. Experimental data can be applied to real-world policy networks, or integrated with real world research. Classroom policy network simulations are used here to accomplish this. This form of proxy simulation offers a plausible way to get more robust results that can be generalizeable to real world policy networks.

This analysis has several important components that, in conjunction, create a unique approach to the question of social structure and policy learning. First, it focuses on relations between actors, not organizations, unlike most policy network research. Second, a substantial amount of social network analysis considers actor's attributes or their relationships, this analysis incorporates both concurrently. Third, learning outcomes are determined using a consistent external measurement (the grade of each actor).[6] Finally, the research design maintains similarity to a real world policy network in ways that other research on social learning does not.

1 Learning Mechanisms

In addition to the potential benefit to learning that social structure may provide in a policy network, this paper also addresses the learning process mechanism that occurs between actors in a policy network. The two primary approaches to this problem can be conceptualized as *information diffusion* versus *collaborative interaction*. Some network measures that examine power and influence are based on the premise that information advantage comes from more exposure to the transmission of information. Information "bits" are simply transferred from one person to another. This is information diffusion. Many discussions of social networks assume that benefits received from the networkare derived from diffusion [46]. This concept presumes that persons who are more central are exposed to additional sources of information that privileges them over those who have less central roles. Further, persons who have access to a wider variety of information act as "brokers" across homogenous groups and bridge the differences (known as "structural holes")

[5] The examination in this chapter is focused on learning performance and process - that is the emphasis is on "puzzling." The relationship between social structure in policy networks and influence/power is the subject of other related research.

[6] Hogset and Barrett argue that some authors depend heavily on proxy reported peer behavior in social learning, which can have significant measurement concerns [28].

between such groups. They have advantages over those whose network ties have greater homophily (similarity) or fewer ties.

A second, different model of network interaction focuses on the development of ideas and learning as a collaborative, implicitly interactive process [11, 54]. The learning and idea production process is likely improved by the extent of collusion, collaboration, and direct interaction that occurs between two people. Learning, knowledge performance, and achievement is improved by person-to-person collaboration and re-examination of information common to all participants, or information that some possess and others do not. Information transmittal is bolstered by consideration between two people - in effect, Hall's "puzzling" process described earlier, where knowledge understanding improves.

Organizational theorists have examined information diffusion extensively. Granovetter's [21] pioneering work demonstrated that "weak" ties exposed one to information they were less likely to know already. This information was more valuable because an actor was exposed to sources of information and points of view not already common to other members of their network. The mechanism underlying this finding was later expanded on by Burt [11] as "structural hole" theory. By definition, weak ties inherently have more structural holes. A person who has ties that are weak invests fewer resources into each of those ties. They can devote their resources (e.g. time, energy, learning potential) to more ties and gain information not already known to them.[7]

Some argue that the transmission of highly complex ideas occurs more with strong ties [25]. In a policy network context, however, most actors already have a high degree of complex policy knowledge necessary for entry into the network. Many of the ideas within a policy network may be moderate or simple additions, or emendations of a body of policy knowledge that is complex, but where additional information is not.

It is less apparent, however, that weak tie theory will be just as applicable in an interactive learning process.[8] In a collaborative context, it's possible that having ties that are more "intense" or frequent may increase the benefit derived from collaboration. Presumably, the benefit derived from intense col-

[7] Burt's structural holes theory is focused less on the strength or weakness of ties, but rather on whether individuals have ties with structural holes in which ego's alters do not interact with each other. The actor can be expected to have greater access to different kinds of information and experience if their connections are to persons who are not simply reinforcing each other's knowledge. Structural holes theory emphasizes the ways in which brokers can improve performance in a variety of ways (e.g. job success, creation of "good ideas," goal achievement) because they bridge gaps between different homogenous subgroups and are thus able to integrate ideas and knowledge from a broad array of sources. Burt notes that successful managers tend to have networks characterized by both structural holes and weak ties. As a general rule, an actor's weak ties are more likely to cross subgroups and have more structural holes - the two conceptual ideas are often collinear.

[8] Assuming that such collaborative learning processes are indeed different from diffusion processes, and that they bring value to the learning process.

laboration or strong ties comes from the development of ideas and learning which can be pursued in much greater depth. This discussion implies four possible models of policy learning, as seen in Figure 1, and implies several hypotheses derived from earlier discussion that we can specifically test.

2 Hypotheses

- H1. Centrality should be associated with an increase in policy learning performance.
  H1a. We can expect the benefit derived from centrality to reduce or flatten as centrality measurements attain high levels.
- H2. The mechanism underlying the learning process will be collaborative.
- H3. Stronger ties or frequent interactions will provide greater benefit.

		Information Transfer Mechanism	
		Collaborative/Interactive	Information Diffusion
Intensity	High	Collaborative/Interactive High intensity (strong ties)	Information Diffusion High intensity (strong ties)
	Low	Collaborative/Interactive Low intensity (weak ties)	Information Diffusion Low intensity (weak ties)

Fig. 1: Four Models of Learning

3 Research Design

Network analysis emphasizes the relationship between actors (dyads), rather than their individual attributes. It is a deeply developed field, particularly in sociology and organizational behavior. It uses matrix algebra for relational analysis and different statistical techniques for inference.[9] The vast majority of research efforts using network methods to examine policy networks have

[9] For an excellent introduction to network analysis see [52, 49].

focused on the organization as the primary unit of analysis [27, 32].[10] One would wish to see network analysis of both individual actors and organizations. Each approach can bring different kinds of clarity for considering political concerns. Organizational data are often easily available but person to person analysis is less common because it is harder to get good data or reliable results. Policy makers and politicians, particularly more powerful or higher level actors, are generally unwilling to fill out questionnaires of sufficient length needed to conduct productive social network analysis. First, they take a lot of time, a commodity in short supply for most people who function in large policy networks. Second, the more "political" actors are legitimately concerned about privacy and publicity issues. Elites may not trust the motivation of the researcher, or the ability of the researcher to keep responses confidential. Further, even if the researcher is trustworthy, there are situations in which a public figure or politician can be legally compelled to publicize their responses. Finally, they may view the kind of survey data they are being asked for as irrelevant, odd, or not directly related to the political, policy, and regulatory questions they are addressing. Even if a researcher can get policy actors to respond to survey requests, it is unpractical to expect complete or "whole" network response. Response rates will generally be quite low. Further, correctly determining the boundaries and participants of a given policy network can be difficult.

This leaves the researcher conducting policy or political research examining network analysis at the individual level with three options. First, they can conduct partial or incomplete network analysis in a policy network, assuming they can get enough responses. This can include ego-network analysis[11] or other techniques. Second, they can use "proxy" network data such as email records and lobbyist statements [33], or political contributions [53]. Third, they can construct some kind of policy network simulation. The focus of this chapter is based on this last option - creating a simulation of a policy network within the classroom.

Outside the policy network arena, social structure has been explicitly linked to learning performance [15]. However, much of this literature fails to account for learning processes in a real-world or policy network context, or neglects to control for important variables (for instance, previous learning performance), or does not measure learning in an objective manner [5]. Scholars have demonstrated strong peer effects on educational performance, showing that if one is surrounded by high performing or high learning alters, then one performs better oneself [14]. Baldwin et al [3] demonstrated that network centrality has a strong relationship to individual learning performance but student interaction in that research was assigned, and not comparable to those of a policy network. If students are assigned to teams, the natural deci-

[10] Less common examples that assess person to person contact include David Lazer's research [33, 34].

[11] This is network analysis that analyzes the personal networks of different respondents but does not link them in a complete network.

sion processes by which actors choose academic or social alters is constricted significantly. Personal interconnections are strongly predicated on trust and homophily (similarity). This research attempts to let as much interaction occur naturally, as it does in a real-world policy network.

A simulation was conducted to determine how the amount (degree) and intensity of collaborative academic interaction between students, using social network measurements of centrality, might affect grades. In the final class week of the semester, 294 students in "Introduction to Political Science" were surveyed on aspects of their class involvement. The academic networks and measured centrality of the class occur as the intervention in a simulated "natural" research design.

To accurately assess the impact of social networks implicitly means that the variable cannot be randomly assigned, nor is it plausible.[12] Choosing whom to interact with in policy/social networks is an important component of real policy networks that should be replicated in simulation. Genuine experimental methods cannot be used. If the centrality variable is randomly manipulated it removes the important self-selection processes that create centrality in social networks. Further, one cannot effectively pre-test for a grade within a class, so this is a post-test research design ([30], 108). However, SAT scores were used as an important control in the statistical analysis.

The questionnaire surveyed students on both their social and academic interactions with student peers. However, only the data on academic interactions were used to create the structural measurements of centrality. The query on social interactions was placed on the questionnaire solely to induce the respondents to clearly differentiate between their social and academic interactions. This works to focus students on the interactions most relevant to their learning performance. Second, it provides an important additional control in "pure" social interaction. For instance, if a student engages in some other social activity with another student, it may provide a higher level of trust in academic interactions.[13] Finally, cleanly delineated academic interactions are an important way in which the classroom interactions are representative and generalizeable as a policy network (whereas social interactions are not).

Students were asked to name other students in the class with whom they had either academic and/or social interactions and to rate the intensity of those interactions as high, medium, or low. High intensity interactions were those that occurred an average of 1-2 times or more per week through the semester. Medium intensity interactions occurred 6-14 times during the entire semester while low intensity interactions occurred 1-5 times. The determination of an "interaction" or episode as academic was made by the student.

[12] For instance, imagine asking policy actors, even students in a policy simulation exercise, to refrain from speaking with anyone but the 1-15 randomly chosen actors that have been assigned to them over a three month period.

[13] Additional analysis looking at the relationship between these variables will be the subject of upcoming research but is not covered in this study.

Other important control variables included motivation to do well in the class, student year status (freshman, junior, etc.), SAT scores, and gender.

The sample was defined by the students enrolled in the class. Although students might conceivably interact academically with students outside of the class on the subject matter, the survey was designed with the assumption that the vast majority of students interact with other students in the class because of their shared experience. Thus, the density of the academic network reported may be slightly depressed because of this boundary definition. Incidental discussion with students confirmed that they rarely discussed academic aspects of the class with students not enrolled in the course.

For the purpose of the experiment, the educational process being tested is not assumed to be affected by the degree of power, competition, or influence exercised by one person over another. Students implicitly understand that to gain an "A" grade they do not have to prevent another student from gaining an "A." Clearly, power and influence play an important role in policy outcomes, and within policy networks. However, in a policy learning process, particularly in earlier stages, it is unclear how information will affect policy outcomes. Stakeholders or coalitions do not yet have a clear position. In effect, the absence of power roles in the classroom functions as a control for the power dynamic in the research.

198 out of 294 potential respondents responded to the survey questionnaire (67% response rate).[14] Obviously, this creates distinct differences between the two versions of the network depending on which way the experiment is bounded. Two versions of data from the class are used in the data analysis. First an "entire class" model was created that used respondent data with non-respondents to create an inferred network, a common technique in network analysis [52]. Data were symmetrized (made equal) in such a way that any interaction was made reciprocal if one node indicated an interaction. Thus, if student M indicated that they had a low interaction with student L but student L did not indicate that they had any interaction with student M then they would both be marked with a low interaction. The symmetrization process produces a default version of interaction that is biased toward a respondent who believes an interaction took place rather than a respondent who might have forgotten the interaction [10]. The weighting or intensity of interactions was not changed. If student M indicated they had a medium interaction with student P and student P indicated they had a low interaction with student M, the data were not modified.

The "respondent only" network also symmetrized data as described above. However, in this model, all non-respondents are eliminated from the data analysis. The "respondent only" version was created to test for correlations

[14] Generally, large lecture classes such as this experience a 15-30% absentee rate on any given day. The response rate of 67% occurred in large part because of a flu outbreak on the survey date. There is no reason to expect bias. Further, network symmetrization (discussed below) can fill in data missing due to non-response or forgetting on the part of one actor in a dyad. This issue is discussed further in Brewer and Webster [10].

with motivation and party affiliation (data that was only available from respondents).

The network represented in the "entire class" model creates a superior representation of the actual network. In Table 3, centrality correlation comparisons are shown for both models of the experiment. The "entire class" model produces slightly more robust statistical results than the respondent only data. However, the differences between the "entire class" versions versus the "respondent only" version are minimal.

4 Findings

Let's consider centrality and learning performance first. Structural centrality measurements for students were derived in Ucinet, a specialized network analysis software [7]. Measurements included degree centrality, structural holes, Freeman degree, eigenvector, and betweenness centrality.[15] Analysis was conducted with versions that weighted the ties by intensity and a non-weighted version in which ties were considered equal. Non-weighted measures of degree centrality and structural holes had the strongest results and were virtually identical.

Degree centrality often has collinear results with structural holes measures, a more sophisticated assessment which accounts for one's connections to others who are not connected to each other [8]. Figure 2 shows the average degree centrality of all students who received that grade. Generally, each successive grade category demonstrates a higher trend towards degree centrality.

Obviously, controls for other cofounders need to be considered. Grade outcomes were regressed on degree centrality measurements with several controls, including dummy variables for gender, student year, and political affiliation. These results are shown in Table 1 below. Model 1 and 2 include a comprehensive set of variables. Unsurprisingly, SAT scores and motivation both were strong predictors of final grade results.[16] Regression results demonstrate strong effects for degree centrality. In order to test for non-linearity, the variables of theoretical interest were transformed by the natural logarithm. This revealed no substantive alterations in the results.

[15] The other centrality measurements make adjustments to account for influence and power mechanisms and can be conceptualized as measuring information hierarchies. There was no statistical relationship for those measurements, other than degree centrality and structural holes. For a technical distinctions between various network centrality measures see [52].

[16] See Appendix 3: 'Descriptive Statistics' for an overview of the breakdown of the class in terms of grade received, affiliation, motivation, and year in school. Appendices are available by contacting the author.

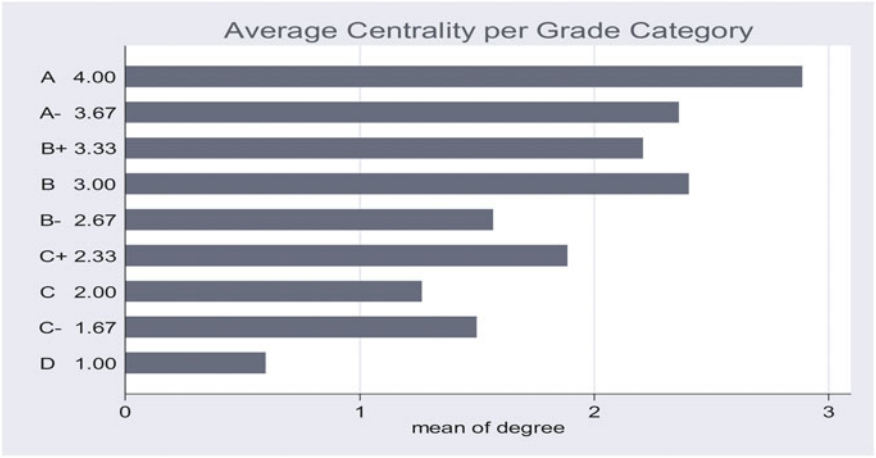


Fig. 2: Average Centrality per Grade Category

Data on political affiliation were included in the survey as a point of interest.[17] Models 3 and 4 showed a statistically significant relationship between political affiliation and grade performance. This relationship receded after SAT scores were introduced. Political affiliation was measured as an ordinal variable running from right to left (libertarian, republican, neutral, democrat, green, socialist/communist). Dummy variables were also created and analyzed for democrat, republican, and "other" affiliations.

Data on student year was analyzed as an interval variable and also in dummy variable form to account for the possibility of year as a categorical or ordinal variable type.

Dummy variables should account for "senioritis" when seniors exhibit less motivation to do well, or account for freshman who have smaller interaction networks than other established students. There was no correlation between the students' years and their grades, nor was there a correlation between student year and their centrality. Is there a limit to the benefit a policy actor may receive from a central position in a policy learning arena? A flattening of beneficial effects occurs once a student attains 6-7 network interactions (see Figure 3). Data were analyzed using quadratic regression procedures in Stata

[17] It is a political science class after all. Recent (and controversial) literature examines links between left political affiliation and intelligence [2, 17].

[18] The dichotomization process is described and discussed in detail later in the chapter. The dichotomized model provided the most robust results. This data transformation process reduces all valued interactions (for instance tie intensity weights) to 1, making all ties dichotomous (valued at 1 or 0).

Table 1: Degree Centrality and Learning Performance (Grade)

Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Degree Centrality	0.07 ** (0.029) 0.18	0.08 *** (0.029) 0.20	0.07 *** (0.028) 0.19	0.08 *** (0.027) 0.20	0.06 ** (0.025) 0.17	0.07 *** (0.025) 0.18
Student Motivation	0.18 *** (0.030) 0.40	0.17 *** (0.031) 0.38	0.14 *** (0.032) 0.30	0.14 *** (0.032) 0.30	0.17 *** (0.029) 0.38	0.16 *** (0.029) 0.37
Political Affiliation (Left)	0.07 (0.044) 0.11		0.09 ** (0.045) 0.14	0.09 * (0.046) 0.13	0.07 (0.043) 0.11	
Republican		0.09 (0.125) 0.05				
Student year	-0.03 (0.054) -0.04		-0.04 (0.057) -0.05			
Soph. or Junior		-0.06 (0.097) -0.04				
Gender	0.001 (0.093) 0.00	-0.01 (0.093) -0.01				
Component Dummy	-0.11 (0.117) -0.07	-0.09 (0.119) -0.06				
SAT score	0.02 *** (0.004) 0.30	0.02 *** (0.004) 0.30			0.02 *** (0.004) 0.33	0.02 *** (0.004) 0.33
Constant	-0.97	-0.82	1.58	1.47	-1.30	-1.08
N	177 (some SAT data not available)	177	198 (all respond.'s)	198	177	177
R Square	0.28	0.27	0.17	0.17	0.29	0.28
F Value	9.31	8.97	9.73	13.17	17.74	22.58

* = Significant at 90% *Variable Descriptors:* Coefficient
*** = Significant at 95% (standard error)
**** = Significant at 99% standardized beta coefficient

All Models above use non-weighted (dichotomized) interaction measurements. [18]
"Models" in this table refers to the different specification of variables for each regression.
The "Component Dummy" variable is explained later in these findings.

to help determine the nature of the interaction.[19] The benefits of adding an additional alter to one's network are substantial when one begins with few

[19] This is the qfitci plot command in Stata. It quadratically regresses the dependent variable on one independent variable. It provides a non-linear representation and estimation of the relationship between the two variables. The quadratic approach can provide a better

alters. From zero to four alters, the benefit of adding an additional contact to one's network provides an average 0.19 increase (on a 4.0 point grade scale), or about a 4.8% increase in grade number. At seven alters and higher, the benefit ebbs appreciably and the fitted values exhibit heteroskedacticity.

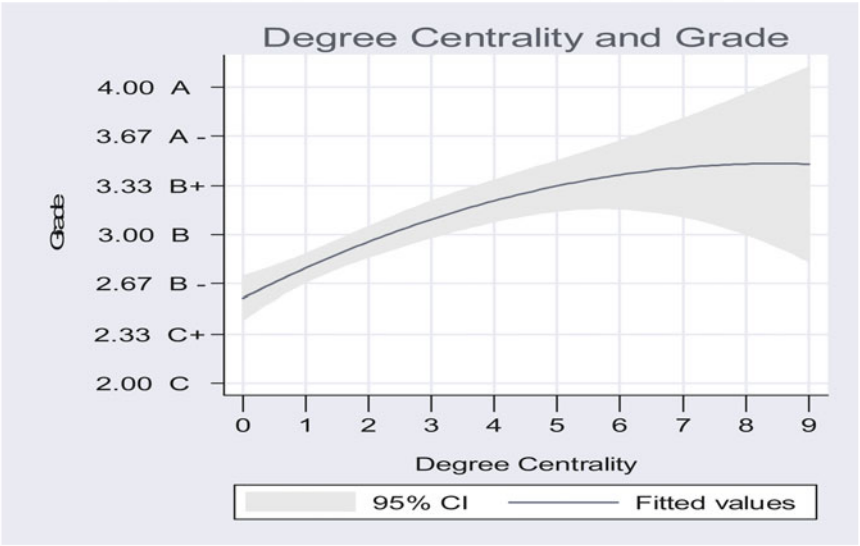


Fig. 3: Degree Centrality and Grade

While one wouldn't expect the linear benefits of adding more alters to one's network to be infinite, one might expect that such a "cap" would occur at a higher level, at around 10-15 alters. However, this counter-intuitive result makes more sense when one considers that students (or policy actors) are negotiating information in a wide variety of different information subcontexts ("mini-networks" if you will). Each of these mini-networks has its own alters and information; an overload of information and/or personalities in any network will begin to negate any benefits from extensive amounts of alters.

fit than a linear regression, depending on whether the relationship is curved or linear, but cannot be used with multiple variables. For further discussion see McLoone [38].

5 Diffusion versus Interaction

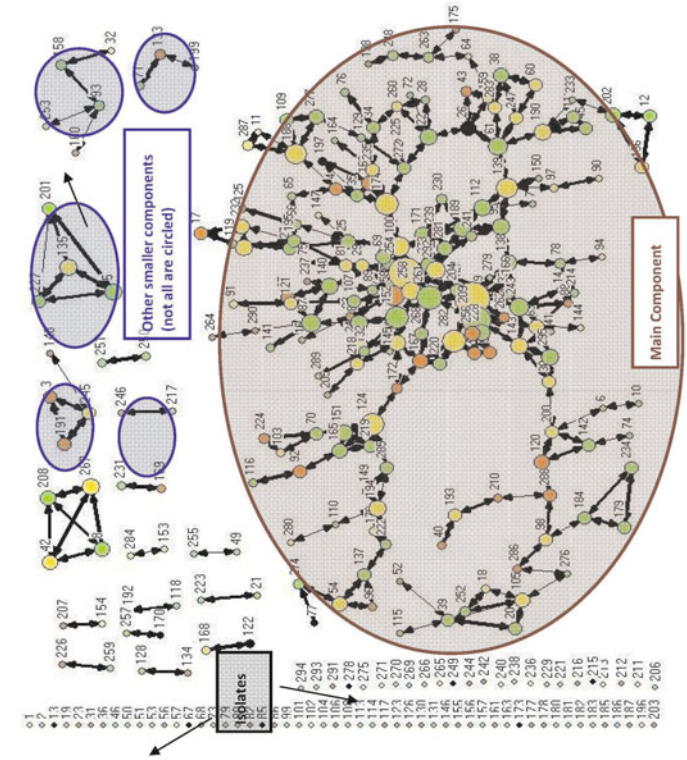
As discussed earlier, the second hypothesis in this work focuses on the nature of information sharing and learning processes. In particular, I contrasted two models of learning, one focused on diffusion and the other on collaborative/interactive mechanisms. One can conceive of diffusion as the passage of information that originates from the professor and/or class readings (or alternately, policy experts and different written policy sources) to the students (other policy actors). The interactive learning effect that occurs with the students is one in which they further diffuse such information to others who missed it the first time around.

Alternately, a collaborative mechanism is informed by a more interactive teaching, learning, and puzzling process that occurs around such pieces of information. Of course, information exposure is part of the process, but increasing information *understanding* may be more important. A situation in which students (as policy actors) truly teach and learn from each other in a process that is qualitatively different from the information they gain from readings and/or the expert (professor).

The network that emerges in this study provides a nice test for the "pure diffusion" concept. In Figure 4 (Entire Class Network), the network diagram shows that there are several separated components (as well as many isolates - those nodes not academically connected to any other node in the network). The very large component in the center (of approximately 100 nodes, or about one third of the class) should have a distinct benefit over the other much smaller components in the class if the learning process is only derived from the diffusion of information. Students in the large node gain the benefit of different information from 99 other alters in that component that are being spread from one actor to another.

Alternately, if an interactive, collaborative learning process is occurring, then a student with ties to 3 other students should benefit equally, regardless of whether they are in a network component of 100 students or 5 students, all other things being equal.

To analyze the question of diffusion versus collaboration, membership in the large central component was set up as a dummy variable for each student. These data were analyzed to see if being a member of the large network component brought any grade benefit when controlling for centrality. Table 2 shows that while the coefficient for component membership did move in the expected direction, the effect was weak and statistically insignificant. In less technical terms, this means that if a person was in a small component of 4 or 5 total persons or in the large central component of over 100 students, the benefit of being academically linked to three other persons in either component would be virtually identical. This result provides strong inferential evidence that the learning process is collaborative, rather than a simple case of informational cascades. If the benefit accrues from the diffusion of information,



Color of Node	Student Grade
Yellow	A
Green	B
Orange	C
Black	D or F

Size of Node	"Degree" Network Rating
Smallest	0
to	to
Largest	20

Line Width	Intensity of Interaction
Thin	Low
Medium	Medium
Thick	High

Fig. 4: Entire Class Network (weighted results)

then there is much more total information to be diffused in the large central component, and actors therein should receive the benefit of that information.

6 Tie Intensity and Learning

Finally, let's consider what kind of ties provide the most benefit to grade. Initially, I expected that interactions which occurred more often ("intensity") would be associated with an increased learning benefit. Ties between students who had only 1 or 2 interactions over the semester would be less beneficial than students who discussed academic/policy issues regularly. This would

Table 2: Learning Performance, Component Effect, and Centrality

Variables	**Model 1**	**Model 2**
Degree Centrality	0.15 *** (0.028) 0.30	0.14 *** (0.035) 0.27
Component (dummy variable)		0.087 (0.111) 0.05
Constant	2.65	2.63
N	294	294
	(entire class)	(entire class)
R Square	0.10	0.097
F Value	30.74	15.66

* = Significant at 90%
*** = Significant at 95%
**** = Significant at 99%

Variable Descriptors: Coefficient (standard error) standardized beta coefficient

All Models above use dichotomized results for interactions.

represent a reasonable exception to Granovetter's weak ties theory because in a learning process one could expect that interacting with another student on a more regular basis would provide a better learning experience. As we shall see, this hypothesis was refuted by the available evidence.

Three different models of centrality were created to test this theory. Each model was tested in full versions of the class (n=294) and the "respondents only" version. Centrality measurements were initially derived using the original weighted data. Recall that students rated their interactions. In the "weighted ties" version, centrality ratings were derived using weighted data. An "intense" interaction was weighted at 3, medium at 2, and weak/minimal at 1. A student who had only two academic interactions but each rated as intense (i.e. 2×3) received a 6 centrality rating. Similarly a respondent with five interactions that were all "minimal" (i.e. 5×1) would receive only a 5 centrality rating. This model incorporates my hypothesis that more "intense" interactions have a greater educational performance value for the respondent.

The second variation in weighting is denoted as "'1' ties purged, '2&3' ties dichotomized." In this variation, all weak intensity ties were discarded. Medium and intense ties were dichotomized (reduced equally to a value of "1"). The assumptions underlying this model were that minimal/weak ties had little value for an increase in educational performance and that remaining ties should remain equal. As is apparent in Table 3, this produced a slightly weaker correlation with reductions in both model fit and proportion of variance explained.

The last variation, "all ties dichotomized," kept all ties. However, this model eliminated all valuations of tie "intensity," reducing all weights to equal measures of 1. This variation produced the best result, particularly

in the 'entire class" testing version. This model produced a beta coefficient of 0.32, accounting for a 10.4% of the variance as well as a strong statistical fit. These results accentuate and reinforce the "weak ties" and "structural holes" theories of Granovetter [21] and Burt [11] discussed earlier.

Table 3: Centrality Models and Learning Performance

Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	Weighted ties	'1' ties purged '2 and 3' ties dichotomized	all ties dichotomized	Weighted ties	'1' ties purged '2 and 3' ties dichotomized	all ties dichotomized
Degree centrality	0.263***	0.251***	0.282***	0.312***	0.286***	0.323***
Constant	1.69	1.69	1.68	2.63	2.65	2.63
N	198 (resp. only)	198 (resp. only)	198 (resp. only)	294 (entire class)	294 (entire class)	294 (entire class)
R Square	0.069	0.063	0.080	0.097	0.082	0.104
F Value	14.590	13.233	16.942	31.497	26.063	33.971

* = Significant at 90%
*** = Significant at 95%
**** = Significant at 99%

Data for centrality above are beta weights, not coefficients.
Model 6 demonstrates strongest r2 and best beta weight: all ties dichotomized

Granovetter argues that resources used for social interaction that are oriented towards weak ties are more useful. They provide information not known by alters in one's network already, and provide information that is not already embedded in one's network. Further, a respondent who uses limited resources for social interactions can gain more by having ties with a low level of "upkeep." Burt's structural holes theory posits that connecting with other actors who are not themselves connected with each other allows the central actor to function as a power and/or information broker while also allowing the central actor to create slightly different "persona" to different portions of their network.

Weak ties and structural holes seem to function in a slightly different context in the educational environment characterized by this experiment. In this situation, solely the amount of ties and/or structural holes appears to demonstrate that actors improve educational performance simply by increasing the amount of interaction amongst as many different actors as possible. The results demonstrate that developing strong ties with few actors may not be as useful as developing ties, even if weak, with as many diverse actors as possible.

This result is substantiated by the dichotomized weighting (all ties weighted equally) which achieves slightly better correlations than the two alternate versions in which strong ties were weighted and weak ties were discarded. Certainly, this reinforces the conceptual idea (described earlier) that the learning process is one of collusion, collaboration, exploration, re-examination, and clarification; all of which increase understanding. If an actor is exposed and

connected to a wide variety of actors, the diversity of the interaction seems to have at least as much value as the intensity of the interaction. If intensity mattered more, then the models that weighted strong ties would have had a higher correlation than the dichotomized version.

7 Discussion

Hanneman ([24], 41) notes that network size affects the density and fragmentation of a network, and also the ease with which information may be transmitted through it. The class network, as measured by its academic interactions, is particularly fragmented (93%), low in its cohesion (6%), with a large average distance between reachable pairs (7.6 ties).[20] Density statistics on the network are presented in Appendix 2.[21] Thus, the present experiment represents a scenario in which the network effect will be conservative. This is important because if causal effects are realized in a non-cohesive, fragmented, and diffuse class of large size (n=294), it is logical that such effects would be greater in a smaller and/or more socially intensive policy network environment. The experiment most likely represents a conservative test of the effect of social networks on learning performance.

There are two key potential concerns for this research. The first involves the assignment problem. While this analysis demonstrates that centrality has a demonstrable effect on learning performance, it may be that it serves only as a proxy for some other variable not considered in the investigation. Perhaps charisma for instance, or some other personal attribute, allows an actor to gain access to more information from others, increase their centrality, and functions as the true underlying mechanism by which an actor improves their learning performance. However, the value of different personal attributes will vary depending on the opinion of the alter. One person's charisma is another person's self-importance or superciliousness. Further, networks reflect a wide variety of input variables. They are shaped by institutional constraints yet also by culture, past relationships, homophily, and other relationship history. Centrality reflects a mixture of cultural, institutional, and actor-centered attributes. Even if centrality functions as an intervening variable, this should be of interest to scholars. One would still wish to know whether and how networks distort other important independent variables, and exactly how networks function as an intervening variable.

In addition, there is considerable evidence that personality attributes have little predictive relationship with network centrality, particularly in professional or strategic interactions. For instance, Novak demonstrates that per-

[20] These percentages utilize the "distance" measurement from Ucinet software [7]. Distance measures are derived from Burt [12].

[21] All appendixes are too long to be included in the published version of this paper. The author is happy to provide them upon request.

sonal characteristics are rarely considered in an actor's strategic or operational networks (2008). His delineation approximates the social/academic delineation used in this survey. Other research in the author's dissertation looking at the relationship between centrality and influence considers personality characteristics via the "big five" personality test [29, 50] and also via dyadic survey response. This analysis demonstrated no statistically robust relationships other than perceived leadership and expertise traits in the dyadic surveys. The perception of other's expertise in information networks is strongly linked to centrality [6]. Other personality traits associated with centrality are the degree of self-monitoring, or conscientious ingratiation into networks [40, 44]. Finally, those who exhibit neurotic tendencies are less likely to become central [31]. Instead, centrality is predicted extensively by homophily or similarity [39], hierarchy [35] and physical proximity [6].[22] Given these considerations, it is reasonable to consider centrality as an exogenous independent variable in this analysis.

The second concern involves external validity. The question of representativeness, or generalizability of the classroom research to real world policy networks may be of concern to the reader. We may recall that policy networks had three unique characteristics: extensive technical knowledge within one policy realm; interaction based on role within network, and focus on policy outcomes.

First, the class has a broadly focused background in one policy realm, political science. The class in this analysis is large (n = 294), diffuse, fragmented, and not initially separated into sub-groups other than discussion sections.[23] The students are linked primarily by their roles as political science majors; over 90% of the students were majors, or planned to be. They have a similarly strong emphasis and interest in the arena of political science in the same way that actors in a policy network will be focused on the issues within their policy realm.

Second, students' focus in class is strongly concentrated on their grade results, their "policy outcome." For the vast majority of the students, their grade is the single most important outcome within the class. In a policy network, actors within coalitions are generally focused on legislative or regulatory outcomes in which they interact with other actors to learn more about the underlying issues that affect those outcomes. Similarly, students will interact with other students to study the results of their policy realm (i.e. a survey-level knowledge of political science) to gain a preferred policy outcome (grade performance).

The third function, that policymakers interact based primarily on their roles within a policy network is the one in which a classroom has less similarity to a policy network. Students are clearly younger, and interact with each

[22] I recommend the excellent annotated bibliography of the antecedents and consequences of centrality by Dan Brass [9].

[23] In this area, this quasi-experiment is different from other research that examines network effects within assigned small cohorts in business school settings [3].

other based on a variety of social identity factors (e.g. social interests, sexual interaction, mutual past-times) much more than actors would within a real policy network. However, this is addressed, at least in part, by the survey distinctions between social and academic interaction. Further, recent work by Druckman and Kam argues strongly that "student subjects are not an inherent problem to experimental research" as long as researchers work to reproduce the correct context in experimental work (as I argue has been done in this analysis). "Moreover, the burden of proof-of student subjects being a problem-should lie with critics rather than experimenters." ([19], 1).

The classroom environment is a useful context in which to illustrate the difference between the two models of learning. A professor lectures to students and distills large amounts of information to them. There is little collaborative or interactive exchanges in this process, particularly if the class format is a large lecture. There may be some interaction, particularly if the professor encourages frequent clarification or question breaks or longer, wide-ranging question and answer sessions as regular components of to their formal presentation.

However, a second learning process is also occurring in the context of such a class. This more informal process occurs between the students in their discussion sections, amongst their peers in study groups, in conversations both in and outside the classroom environment, and in some cases via electronic communications. These less formal interactions are characterized by collaborative dialogue and information exchange that is bi-directional and could inherently involve a greater degree of information development between tied parties. It is this portion of the learning process that has inherent similarities to the learning processes that occur in policy networks.

Given these considerations, there is certainly a strong degree of external validity from the classroom results that can be considered relevant to policy learning. The simulation allows for significantly more control than natural settings. Another important consideration is that policy learning is extremely difficult to operationalize in a real world setting whereas using a grade proxy in the class provides a consistent dependent variable.

This work is one portion of multi-state research and any loss in generalizability is made up for by the gains in data that would be impossible to get in the field. Further, this research is meant to be considered within a broader analysis incorporating triangulation [51]. It is meant to support already existing qualitative research in the area of policy learning [23, 27, 48], and to work in conjunction with my other qualitative and quantitative dissertation analysis.

8 Conclusion

Members of a policy network who link to more actors within a policy network without devoting extensive levels of time or frequency to those relationships can expect benefits. They can anticipate greater access to a more diverse set of information, and presumably, more diverse methods of interpreting and understanding such information by linking. Clearly knowledge intensive work and learning requires extensive person-to-person contact which goes beyond the standard process of informational processing [16]. This analysis demonstrates that an informational environment in which one links to their less embedded network alters has a strong impact within collaborative-based learning processes. Five results accrue from this analysis:

1. There is a strong association between the amount of one's network ties and learning performance in a social learning environment. (Hypothesis 1 - confirmed).
2. The benefit of such ties is stronger at lower levels, and is capped or reduced at higher levels. (Hypothesis 1a - confirmed).
3. The learning process is more than simple diffusion or information transmission. Rather, it is a process that is most likely enhanced by collaborative, interactive discussions which allow participants to understand the information in more complex and nuanced ways. (Hypothesis 2 - confirmed).
4. There is no benefit to having intense or frequent ties among network actors for policy learning. It is more beneficial to have "weak ties" to alters who are less connected to one's own network. (Hypothesis 3 - disproved)
5. This simulation has relevant application to real world policy networks.

Network centrality has a strong relationship with increased learning performance measured by grades. This result demonstrates that collaborative, interactive learning processes improve the learning experience. There is some evidence that suggests increasing the amount of such ties/processes is of greater or equal value than increasing the intensity of such ties. Further, the benefit of increasing such ties is stronger amongst actors with fewer alters, and is limited or capped at higher levels.

An unexpected result in this analysis is the relative strength and robustness of the weak tie impact when one might expect more robust results when strong ties are emphasized. There are two possible explanations for this result. First, the learning process may be improved by increasing different kinds of collaboration (i.e. more collaborations with a variety of different students) rather than simply increasing the intensity of the collaborative learning process. Students and policy actors who have many connections, even weak, may be gaining exposure to different kinds of collaborative experience and a wider variety of information.

A second explanation is that the learning process is improved by access to structural holes and that a larger information flow is actually occurring

than is presumed in this research. Diffusion and flow of information are occurring more than is indicated. If these circumstances were present one would expect statistically significant results for alternate centrality measurements such as eigenvector or betweenness centrality. However, those measures had no statistical association with increased grades.

Lastly, the differentiation between intense and mild levels of interaction reported by the respondents may actually be quite minimal. Future research should investigate the effect of supervised or managed highly intensive interaction that is directed towards increasing academic interaction, for instance, in formalized agenda-driven study groups. While the classroom environment is clearly not a direct proxy for policy networks, one can reasonably generalize these results. The class environment has its own policy realm, with actors focused on a common policy outcome. Further, this research supports other qualitative accounts of social learning in policy networks. One of the primary concerns facing advocates of social network theory concerns causal claims [18]. The fact that centrality measures have an independent relationship with learning performance, while SAT scores are controlled for, reinforces the case for a social structure effect on policy learning. The analysis demonstrates that what many professors have intuitively emphasized to their students for years also applies to learning in policy networks: work together on homework, exchange ideas, form study groups, ask questions of each other, and get other perspectives to increase understanding.

References

1. Adam, Silke and Hanspeter Kriesi. 2007. "The Network Approach." In Theories of the Policy Process, ed. Paul Sabatier: 129-154. Cambridge, MA: Westview Press.
2. Alford, John R., Carolyn L. Funk, and John R. Hibbing. 2005. Are Political Orientations Genetically Transmitted? American Political Science Review 99, no. 02: 153-167. http://www.journals.cambridge.org/action/displayAbstract?fromPage=online&aid=307693&fulltextType=RA&fileId=S0003055405051579. Accessed June 08, 2005.
3. Baldwin, Timothy T., Michael D. Bedell, and Jonathan L. Johnson. 1997. The Social Fabric of a Team-Based M.B.A. Program: Network Effects on Student Satisfaction and Performance. The Academy of Management Journal 40, no. 6: 1369-1397.
4. Bonacich, Phillip. 1987. Power and Centrality: A Family of Measures. American Journal of Sociology 92, no. 5: 1170-1182.
5. Borgatti, Stephen P and Inga Carboni. 2007. Measuring Individual Knowledge in Organizations. Organizational Research Methods 10, no. 3: 449-462.
6. Borgatti, Stephen P and Rob Cross. 2003. A Relational View of Information Seeking and Learning in Social Networks. Management Science 49, no. 4: 432. http://proquest.umi.com/pqdweb?did=342679911&Fmt=7&clientId=3740&RQT=309&VName=PQD.
7. Borgatti, Stephen P, Martin G. Everett, and Linton.C. Freeman. 2002. Ucinet 6 for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.
8. Borgatti, Stephen P. 1997. Structural Holes: Unpacking Burt's Redundancy Measures. Connections 20, no. 1: 35-38. http://www.analytictech.com/connections/index.htm.

9. Brass, Daniel J. 2003. "Social Networks in Organizations: Antecedents and Consequences." http://gatton.uky.edu/Faculty/brass/ConsequencesofSocialNetworks.doc.
10. Brewer, Devon D. and Cynthia M. Webster. 2000. Forgetting of friends and its effects on measuring friendship networks. Social Networks 21, no. 4: 361-373. http://www.sciencedirect.com/science/article/B6VD1-3YHG8BC-3/2/741e2a73f973cc7098d45aad15bb9537.
11. Burt, Ron. 1992. Structural Holes: The Social Structure of Competition. Cambridge: Harvard University Press.
12. Burt, Ronald S. 1976. Positions in Networks. Social Forces 55: 93-122.
13. Busenberg, George. 2001. Learning in Organizations and Public Policy. Journal of Public Policy 21, no. 2: 173-189.
14. Carrell, Scott E., Richard L. Fullerton, and James E. West. 2008. Does Your Cohort Matter? Measuring Peer Effects in College Achievement. National Bureau of Economic Research Working Paper Series No. 14032. http://www.nber.org/papers/w14032.
15. Choi, Syngjoo, Douglas Gale, and Shachar Kariv. Social Learning in Networks: A Quantal Response Equilibrium Analysis of Experimental Data. http://socrates.berkeley.edu/ kariv/CGK_I.pdf. Accessed July, 2008.
16. Cross, Rob, Andrew Parker, Laurence Prusak, and Stephen P. Borgatti. 2001. Knowing What We Know: Supporting Knowledge Creation and Sharing in Social Networks. Organizational Dynamics 30, no. 2: 100-120.
17. Deary, Ian J., G. David Batty, and Catharine R. Gale. 2008. Bright Children Become Enlightened Adults. Psychological Science, no. 1: 1-6. http://dx.doi.org/10.1111/j.1467-9280.2008.02036.x.
18. Doreian, Patrick. 2001. Causality in Social Network Analysis. Sociological Methods and Research 30, no. 1: 81-114.
19. Druckman, James N. and Cindy D. Kam. 2009. Students as Experimental Participants: A Defense of the 'Narrow Data Base'. SSRN eLibrary. http://ssrn.com/paper=1498843.
20. Freeman, Linton C. 1979. Centrality in Social Networks: Conceptual Clarification. Social Networks 1: 213-239.
21. Granovetter, Mark. S. 1973. The Strength of Weak Ties. American Journal of Sociology 6: 1360-1380.
22. Greener, Ian. 2001. Social Learning and Macroeconomic Policy in Britain. Journal of Public Policy 21, no. 2: 133-152.
23. Hall, Peter. 1993. Policy Paradigms, Social Learning, and The State: The Case of Economic Policymaking in Britain. Comparative Politics 25, no. 3: 275-296.
24. Hanneman, Robert A. Introduction to Social Network Methods. http://faculty.ucr.edu/ hanneman/soc157/nettext.pdf. 2003.
25. Hansen, Morten T. 1999. The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge Across Organization Subunits. Administrative Science Quarterly 44, no. 1: 82-111.
26. Heclo, Hugh. 1978. "Issue Networks and the Executive Establishment." In The New American Political System, ed. Anthony King: 87-124. Washington D.C.: American Enterprise Institute.
27. Heinz, John P., Edward O. Laumann, Robert L. Nelson, and Robert H. Salisbury. 1993. The Hollow Core: Private Interests in National Policymaking. Cambridge, MA: Harvard University Press.
28. Hogset, Heidi and Christopher B. Barrett. Social Learning, Social Influence and Projection Bias: A Caution on Inferences Based on Proxy-Reporting of Peer Behavior. http://ssrn.com/paper=1141870. Accessed July 25, 2008.
29. John, Oliver P. and Sanjay Srivastava. 1999. "The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives." In Handbook of Personality: Theory and Research, ed. Oliver P. John and Lawrence A. Pervin. New York: Guilford.
30. Kazdin, Alan. 1998. Research Design in Clinical Psychology. Boston: Allyn and Bacon.

31. Klein, Katherine J., Beng-Chong Lim, Jessica L. Saltz, and David M. Mayer. 2004. How Do They Get There? An Examination Of The Antecedents Of Centrality In Team Networks. Academy of Management Journal 47, no. 6: 952-963.
32. Knoke, David, Franz Urban Pappi, Jeffrey Broadbent, and Yutaka Tsujinaka. 1996. Comparing Policy Networks: Labor Politics in the U.S., Germany, and Japan. Cambridge: Cambridge University Press.
33. Lazer, David, Daniel Carpenter, and Kevin Esterling. 2004. Friends, Brokers, and Transitivity: Who Informs Who in Washington Politics. Journal of Politics 66, no. 1: 224-246.
34. Lazer, David, Kevin Esterling, and Daniel Carpenter. 1998. The Strength of Weak Ties in Lobbying Networks: Evidence from Health Care Politics. Journal of Theoretical Politics 10: 417-444.
35. Lincoln, James R. and Jon Miller. 1979. Work and Friendship Ties in Organizations: A Comparative Analysis of Relational Networks. Administrative Science Quarterly 25: 181-199.
36. May, Peter. 1992. Policy Learning and Failure. Journal of Public Policy 12, no. 4: 331-354.
37. McCool, Daniel C., ed. 1995. Public Policy Theories, Models, and Concepts: An Anthology. Englewood Cliffs, NJ: Prentice Hall.
38. McLoone, Jon. 2008. "Linear and Quadratic Curve Fitting Practice." The Wolfram Demonstrations Project. http://demonstrations.wolfram.com/LinearAndQuadraticCurveFittingPractice/.
39. McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. Annual Review of Sociology 27: 415-444.
40. Mehra, Ajay, Martin Kilduff, and Daniel J. Brass. 2001. The Social Networks of High and Low Self-Monitors: Implications for Workplace Performance. Administrative Science Quarterly 46, no. 1: 121-146.
41. Meseguer, Covadonga. 2005. Policy Learning, Policy Diffusion, and the Making of a New Order. The ANNALS of the American Academy of Political and Social Science 598, no. 1: 67-82.
42. Nilsson, Máns. 2005. Learning, Frames, And Environmental Policy Integration: The Case Of Swedish Energy Policy. Environment and Planning C: Government & Policy 23, no. 2: 207-226.
43. Novak, Dan. 2008. Leadership of Organizational Networks: An Exploration of the Relationship between Leadership and Social Networks in Organizations. PhD Dissertation, Regent University.
44. Oh, Hongseok and Martin Kilduff. 2008. The Ripple Effect of Personality on Social structure: Self-monitoring origins of network brokerage. Journal of Applied Psychology (forthcoming).
45. Powell, Walter W., Kenneth W. Koput, and Laurel Smith-Doerr. 1996. Interorganizational Collaboration and the Locus of Innovation: Networks of Learning in Biotechnology. Administrative Science Quarterly 41, no. 1: 116-145.
46. Rogers, Everett M. 2003. Diffusion of Innovations. New York: Free Press.
47. Sabatier, Paul, ed. 1999. Theories of the Policy Process. Cambridge, MA: Westview Press.
48. Sabatier, Paul and Hank Jenkins-Smith, eds. 1993. Policy Change and Learning: An Advocacy Coalition Approach. Boulder: Westview Press.
49. Scott, John. 2000. Social Network Analysis: A Handbook. Thousand Oaks, CA: Sage Publications.
50. Srivastava, Sanjay. Measuring the Big Five Personality Factors. http://www.uoregon.edu/ sanjay/bigfive.html. Accessed July, 2008.
51. Tarrow, Sidney. 1995. Review: Bridging the Quantitative-Qualitative Divide in Political Science. Review of Designing Social Inquiry: Scientific Inference in Qualitative Research by King, Gary; Keohane, Robert O.; Verba, Sidney. The American Political Science Review 89, no. 2: 471-474. http://www.jstor.org/stable/2082444.

52. Wasserman, Stanley and Katherine Faust. 1994. Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.
53. Waugh, Andrew Scott. 2008. Campaign Contributions and Legislative Cosponsorships in the House of Representatives, 1980-2006. Paper presented at Southwestern Political Science Association. http://www.andrewwaugh.com/waugh_swpsa2008_cosponpaper.pdf.
54. Wenger, Etienne. 1998. Communities of Practice: Learning, Meaning and Identity. New York: Cambridge University Press.
55. Zuckerman, Alan S., ed. 2005. The Social Logic of Politics: Personal Networks as Contexts for Political Behavior.Philadelphia: Temple University Press.

Part IV
Social Media

A Journey to the Core of the Blogosphere

Darko Obradović and Stephan Baumann

Abstract Blogs are popular communication instruments in today's web and altogether, they form the so-called blogosphere. This blogosphere has repeatedly been subject to structural analyses, and one of the findings has been the discovery of the A-List phenomenon, a cohesive group of influential blogs in the center of the blogosphere, whose exact identification remained an open issue. We use six language-specific subsets of the blogosphere, for which we aggregated the blogroll-based networks. We adapt core theory to analyse and compare the cohesion in these six data sets, and provide a new robust and scalable method for the identification of core-periphery structures in blog networks, which can contribute to identify A-List blogs more reliably. We demonstrate the effectiveness and robustness by comparing the results to random networks and by cross-checking the six language data sets.

1 Introduction

1.1 The Blogosphere

Blogs (weblogs) are an interesting phenomenon that arised with the Web 2.0. Defined as "dynamic Internet pages containing articles in reverse chronological order" [2], they can be utilised for various purposes by their authors, as

Darko Obradović
German Research Center for AI (DFKI) & University of Kaiserslautern, Germany
e-mail: darko.obradovic@dfki.uni-kl.de

Stephan Baumann
German Research Center for AI (DFKI), Berlin, Germany
e-mail: stephan.baumann@dfki.de

private online diaries, for so-called *citizen journalism* or even for corporate information delivery, just to name a few categories.

Blogs offer rich possibilities for interaction. Authors can include not only textual and multimedia content, but also link to original content, refer to articles in other blogs, maintain a collection of links to other blogs, or let visitors post comments to their articles, which often also contain links. Thus blogs can and do link to each other. The resulting network of blogs forms the *blogosphere*, which attracted many researchers eagerly analysing its structure and dynamics. This is usually done quantitatively with methods and tools from the field of *Social Network Analysis* (SNA), or qualitatively with visualisations and field studies.

One of the findings is the discovery of the *A-List* blogs [2, 5, 9, 11], described by [8] as "those that are most widely read, cited in the mass media, and receive the most inbound links from other blogs". The studies have revealed that these blogs also heavily link among each other, but rarely to the rest of the blogosphere. This rest is often referred to as the *long tail* [13] and consists of millions of blogs that are only partially indexed (Deep Web Phenomenon).

In summary, there is a broad consensus about three attributes that characterise the group of A-List blogs, to which we will refer a number of times in this chapter:

1. A-List blogs are often linked from the long tail
2. A-List blogs often link to each other
3. A-List blogs rarely link to the long tail

For languages other than English, there exist separate blogospheres, as a kind of sub-communities, that are more cohesively interlinked amongst each other than the blogosphere as a whole is. This resulted in language-specific studies, e.g. a quantitative and structural analysis of the spanish blogosphere [14] or the german one [10]. This phenomenon can be a good opportunity for cross-checking and comparing findings.

1.2 Rationale

There exists a strong demand for rankings in the blogosphere, serving as a motivation for blog authors on the one side, and as a filter for blog readers on the other side. Thus, many services attempt to rank the top blogs in some way, e.g. Technorati[1] is the most prominent player in this field. When looking through these lists, one will usually find roughly the same set of blogs in a very different order [8], although they are all based on algorithms counting inbound links. The discrepancy of the algorithms results from different indexes, timeframes and weights.

[1] `http://www.technorati.com/`

While all these algorithms focus mostly on the first A-List characteristic, namely a large number of inbound links, we decided to have a closer look at the effect of the other two characteristics. These two, and especially the second one, the intensive linking among A-List blogs, demand a certain level of cohesion among A-List blogs, which has been mostly ignored up to date. This seems to be well-suited for further quantitative analyses concerning cohesion. While cliques are too restrictive for real-world networks, the relaxed concept of cores [12] appears highly promising in this context.

In this chapter, we want to find out, if the second characteristic of A-List blogs holds in reality, how it can be best detected, and what its manifestation looks like in the blogosphere. These results could then be used to refine the detection of A-List blogs significantly.

The rest of the chapter is organised as follows. In Section 2 we present the aggregation of our test data sets, in Section 3 we then discuss existing core-theories, compare them to other cohesiveness measures and derive a model that fits our scenario the best. Later on in Section 4, we use this model to extensively analyse and compare our data sets, and then, in Section 5, look into some possibilities of using these results to reliably identify A-List blogs. Finally, we conclude this chapter in Section 6.

2 Data Set Acquisition

2.1 Blog Seeds

In order to find a large set of popular blogs, we need a starting point, i.e. a seed list of some popular blogs. First of all, we decide to create six different data sets according to their language. As mentioned in Section 1, blogs of different languages are small blogospheres on their own, and thus we will be able to cross-check our results between these data sets. We have chosen six european languages, english, german, french, spanish, italian and portuguese, which we can all understand, so that the interpretation of the results is assured.

We start with top 100 blogs from existing ranking services, ignoring their positions in these lists. For english blogs, we use the market leader Technorati. For german blogs, we use the German Blogcharts[2], a Technorati-based list. For spanish, french, italian and portuguese blogs, Alianzo[3] provides good lists by language, which we use in this case.

[2] http://www.deutscheblogcharts.de/

[3] http://www.alianzo.com/

2.2 Using the Blogroll

The blogroll of a blog is an explicit list of recommendations of other blogs by the author. We choose to use these links instead of references from articles or comments for a number of reasons.

From a social network point of view, an explicit recommendation link by the blog author(s) is much more expressive and better to be interpreted than an arbitrary reference whose semantics is unknown without a reliable link analysis. Additionally, there are no weights and no timeframes to be considered. All entries are equal, and if an author decides not to recommend a blog anymore, he should remove the corresponding link from his blogroll. Of course, in certain cases the blogroll might be forgotten and outdated, but we expect this to be rather an exception than the rule in a popular blog.

Nevertheless there are some doubts about the expressiveness of blogroll links in the blogging community to be aware of. Some people argue that bloggers might use their blogroll more for identity management than for real recommendations, i.e. they choose the links in order to communicate a desired impression they want others to have about them. Psychologically, this is neither new nor implausible, but we decide to stick to the objective facts here, keeping this possibility in mind.

2.3 Crawling Blogroll Links

We implemented a set of scripts to find the entries from the individual blogrolls, if present. We encountered three pitfalls in this task. First, we had to develop a sufficiently good heuristic for locating the blogroll entries, as their inclusion on the blog page is not standardised in a way we could rely on.

The second pitfall is the existence of multiple URLs for one blog. We check every single blogroll entry with an HTTP request in order to not insert blog links to synonymous or redirected URLs another time into our database. This would cause a split of one blog into two separate nodes and thus distort our network and our results. This is a common problem, e.g. Technorati often ranks a blog multiple times and thus leads to biased results.

The last pitfall is the reachability of a blog. Blogs that are not reachable during our crawl, either because of network timeouts or because they prohibit spiders, are ignored with respect to their own blogroll links, but remain in the data set and can be recommended by other blogs of course.

2.4 Crawling the Data Sets

Starting from the seeds and their blogroll links, we iteratively include new blogs, following a variant of *snowball sampling* [6]. The most often referenced URLs are checked and included, if they are indeed blogs written in the matching language. To decide, whether an URL hosts a blog, we check it via the Technorati API. This works very well for popular blogs, as they are usually indexed. Small blogs from the long tail might remain undetected though. This is less of a problem for our core analyses, as only popular blogs with a certain number of inbound links are candidates for inclusion anyway. The language is detected by counting stop-words in the blog articles. Thanks to the usually rich textual content of blogs, this works very reliably as well.

The data for the english, german, french and spanish blogs has been collected throughout September to December 2008, and the data for the italian and portuguese blogs throughout August to October 2009. All resulting networks are available as Pajek files on the author's homepage[4].

Table 1 lists the relevant interconnectivity measures of the seed lists, i.e. the number of links, the density and the number of isolated blogs with respect to weak connectivity. Notably, all metrics indicate a good interconnection in the language-specific seeds, with the exception of the italian and portuguese ones. The seed lists with 49 and 100 blogs were too small, but we will see later that nevertheless these seeds were sufficient to deliver good data sets with our iterative extension.

Table 1: Seed Network Comparison

	en	**es**	**de**	**fr**	**it**	**pt**
blogs	100	100	100	100	49	100
links	183	376	289	181	52	58
density (in %)	1.85	3.80	2.92	1.83	2.21	0.59
average degree	3.7	7.5	5.7	3.6	2.1	1.2
isolated blogs	31	10	11	18	25	50

Table 2 lists our final data sets after the iterative extensions. As density is hard to compare in networks of different sizes, we additionally list the average total degrees of the sets. Noticeably, we end up with very well interconnected sets of blogs. As expected in blog networks, the degree distributions for both, incoming and outgoing edges, resemble power laws in all six networks [13]. Due to the nature of our extension, we also list the minimum in-degree a candidate URL must have had in order to be checked and eventually included. This value will be of importance later on, as it is a decisive value for the core analyses.

[4] http://www.dfki.uni-kl.de/~obradovic/data/

Table 2: Extended Network Comparison

	en	es	de	fr	it	pt
blogs	8,401	5,373	1,837	3,402	2,773	3,776
links	452,234	104,241	24,065	90,546	75,421	93,770
density (in %)	0.64	0.36	0.71	0.78	0.98	0.66
avgerage degree	107.7	38.8	26.2	53.2	54.4	49.7
isolated blogs	3	0	2	7	11	25
min. in-degree	12	8	5	8	7	9

3 Core Model

3.1 Notations

First of all, we summarise the terms and notations we will adhere to in the following sections. A directed graph G is defined as $G = (V, E)$ with V being the set of nodes, and $E = (V \times V)$ being the set of directed edges of the graph. $n = |V|$ is the number of nodes, and $m = |E|$ is the number of edges in the graph.

Given a node v, the function $indeg(v)$ returns the in-degree of v, i.e. the number of incoming edges. The function $succ(v)$ returns the set of all successor nodes of v, and the function $pre(v)$ returns the set of all predecessor nodes of v.

3.2 Existing Core Models

The intuitive notion of a core has been initially formalised by Seidman [12]. He defines k-cores in an undirected network as connected components that contain only nodes of a minimum degree of k. Thus each node has a maximum k so that it is part of a k-core, but not part of a $k+1$ core. This results in a *core collapse sequence* of the network, where each node is assigned a certain degree of *coreness*. A corresponding algorithm can be implemented in very good polynomial runtime complexity, as we will show in Section 3.4. However, this model has not yet been properly transferred to directed graphs.

Doreian and Woodard [7] provide a good comparison of the core model with other measures of cohesion like cliques, n-cliques, n-clans, k-plexes or density (see [7] p. 269f). In summary, the main advantage of k-cores for the identification of cohesive subgroups is the fact that it partitions the graph in a discrete and iterative manner, whose results are relatively easy to interpret, compared to long overlapping lists of cliques and the like. Additionally, blogroll links have no real meaning for transitivity, which favours the core

model for our approach, opposed to k-cliques and the like, which are based on distances.

The similar task of defining a core/periphery structure of a network has been engaged by Borgatti and Everett [4] using block modelling. In a first step they re-arranged the adjacency matrix in order to identify a core-region and a periphery region. The model also works for directed and valued graphs. However, a large drawback is the use of the adjacency matrix, and a genetic algorithm for finding a re-arrangement with a statistically good fit out of the $m!$ possibilities. This makes it very expensive to apply this model to large graphs.

In summary, we prefer to use Seidman's core definition for our study, as it is more efficient to calculate, and more intuitive as well.

3.3 Core Models for Directed Graphs

Seidman's definition of cores can be intuitively extended to directed graphs, what we need to do in order to apply it in our blogroll networks. Adhering to the terminology for directed graphs and following some initial ideas from [7], we see five options to define a k-core in a directed network:

weak k-core:	when each node has at least k links of any kind to the rest of the core
strong k-core:	when each node has at least k strong connections to the rest of the core, i.e. reciprocal links
k-in-core:	when each node has at least k incoming links from the rest of the core
k-out-core:	when each node has at least k outgoing links to the rest of the core
balanced k-core:	when each node has at least k incoming and k outgoing links to the rest of the core

Options number one and four are uninteresting, because they allow blogs to be part of the core, that have no inbound links at all. Thus, anyone could make himself part of such a core easily, without any external legitimation. As blogs do not have to maintain a blogroll in order to be important, could have been temporarily unreachable during our data acquisition, or have not been covered by our blogroll detection heuristics, requiring outgoing links does not make sense here. Consequently, options number two and five are also not applicable in our case.

When remembering the characteristics of an A-List set, it is obvious that incoming links are the decisive element, and that we consequently will focus on option number three, namely *k-in-cores*. For each core member, it assures a certain authority by the rest of the core. This is consistent to the requirements of Borgatti and Everett's core/periphery variant for directed graphs.

3.4 The In-core Algorithm

We present a possible procedure for determining the in-core values of all nodes in a graph, and .discuss the runtime complexity afterwards. Starting with $k = 1$ and all nodes marked as non-collapsed, we iteratively repeat the following steps.

1. for each non-collapsed node, check if it has at least k non-collapsed predecessors; if not, let it collapse with an in-core value of $k - 1$
2. for each node v collapsed in this iteration, recursively repeat the check of the previous step for all nodes in $succ(v)$
3. if there were no more collapses in the last step, either terminate the algorithm in case that all nodes have collapsed, or proceed to the next iteration with $k = k + 1$

First of all we take a look at the maximum possible value for k in a graph with a given number of edges m. To form a k-in-core with n_k nodes, we need at least $n_k \cdot k$ directed edges, with $n_k > k$ when operating on a simple graph. Due to this last condition, in order to maximise k to k_{max}, we will use a maximally connected component with $k_{max} + 1$ nodes. In consequence, with a given number of m edges, we can reach at most $k_{max} = \lfloor \sqrt{m} \rfloor - 1$, which is thus the maximum number of iterations for the algorithm described above.

In each iteration, step 1 requires to check at most n nodes, looking at their on average $\frac{m}{n}$ predecessors each. This results in costs of at most m per iteration. Independently from the loop, step 2 is executed for at most n nodes throughout the algorithm, as each node collapses exactly once. With an average of $\frac{m}{n}$ successors to be checked again, and each check costing $\frac{m}{n}$ for checking their predecessors, step 2 costs at most $n \cdot \frac{m}{n} \cdot \frac{m}{n}$.

Step 3 can be performed during step 1 of the next iteration, so the total maximum cost for executing the algorithm is within $O(m^{1.5} + \frac{m^2}{n})$. As we only store in-core values for each node, the space complexity remains linear.

4 Core Analysis

4.1 Comparison to Random Networks

In a first step, we compare the core collapse sequence of each data set with the one from a randomly generated network that has the same degree distribution with respect to a node's total degree. This means, that for every node from the original network, there exists a node in the random network with the same in-degree and out-degree. The random networks are generated according to the *configuration model* [3]. This method detects anomalies in the original network, that implicate the existence of a cohesive core group, which does

not exist in random networks to that extent. This method has its origin in a paper by Watts and Strogatz from 1998 [15], and is a popular method for analyses in SNA.

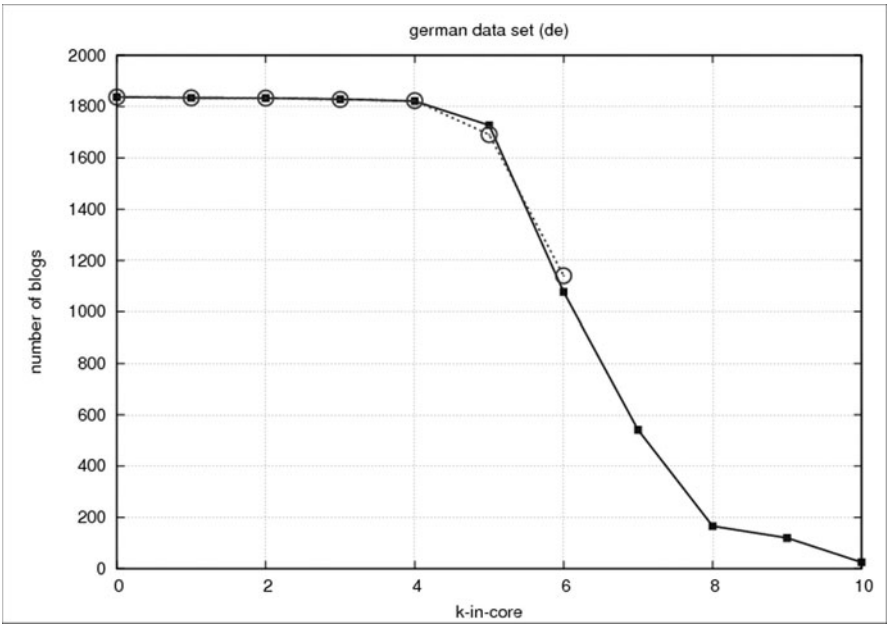


Fig. 1: In-cores for german blogs

The plots in Figure 1 to Figure 6 illustrate the in-core structure of each data set. For each k on the x-axis, the y-axis indicates the number of blogs that are part of this k-in-core. Each plot contains the sizes of the k-in-cores of the original blog data set, marked by filled square points that are joined by straight lines, as well as the sizes of the k-in-cores of the random network, marked by circles that are joined by dotted lines.

In all six cases, we can clearly see that the original data sets tend to contain in-cores with a higher k than expected from the network degree distribution. This means that, beyond the preferential attachment model [1] with its head of high in-degree-nodes and its long tail, these blog data sets have an unexpected tendency towards core-centralisation. This highly conforms to the second A-List characteristic as defined in Section 1.

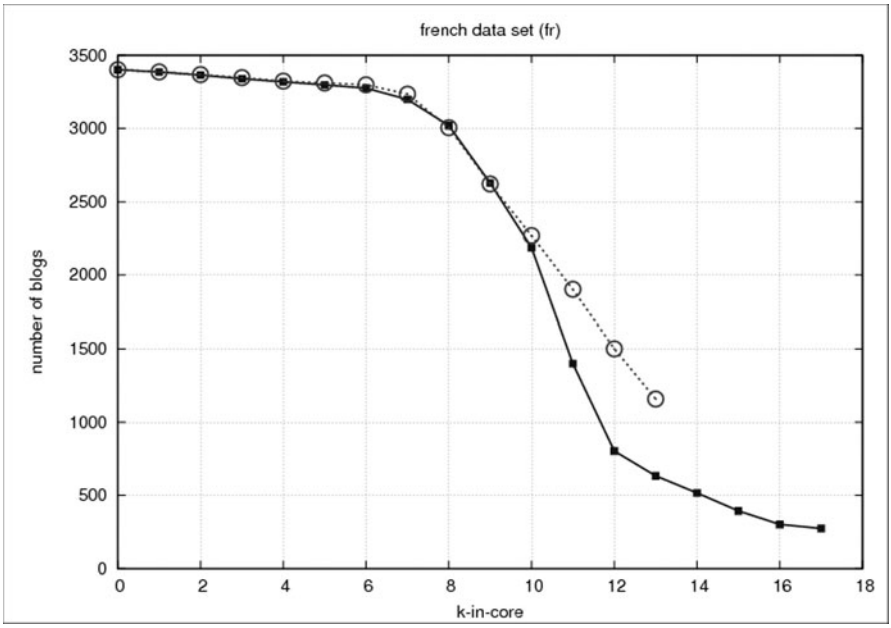


Fig. 2: In-cores for french blogs

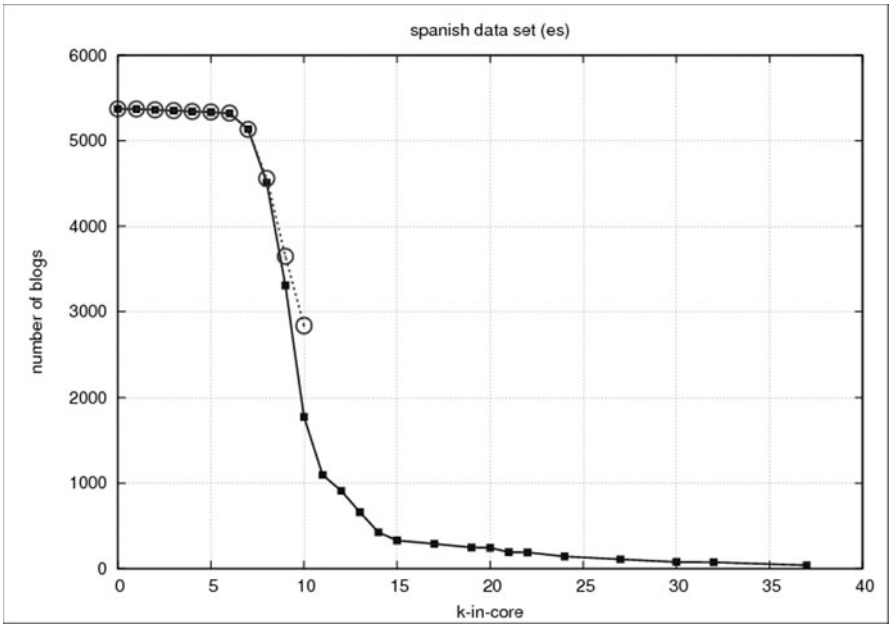


Fig. 3: In-cores for spanish blogs

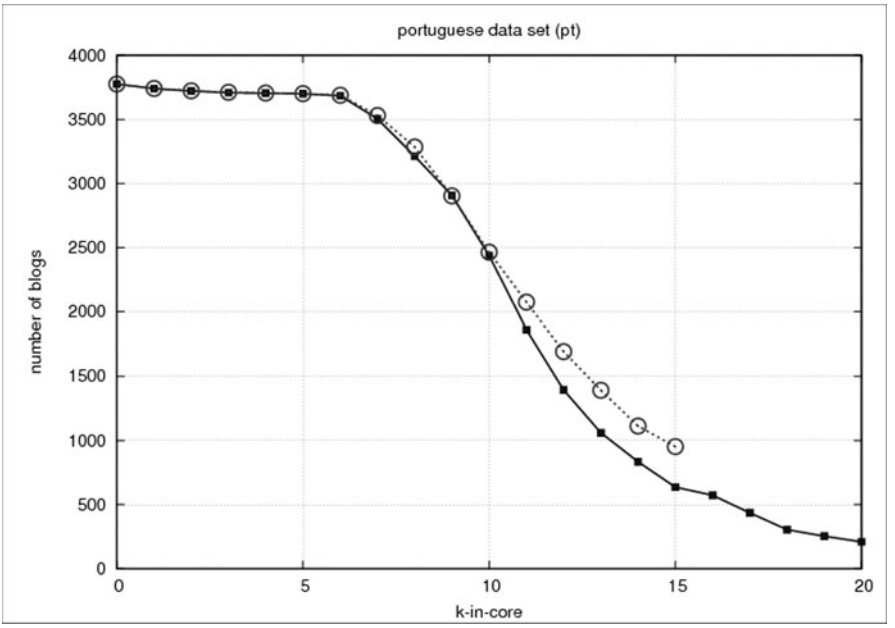


Fig. 4: In-cores for portuguese blogs

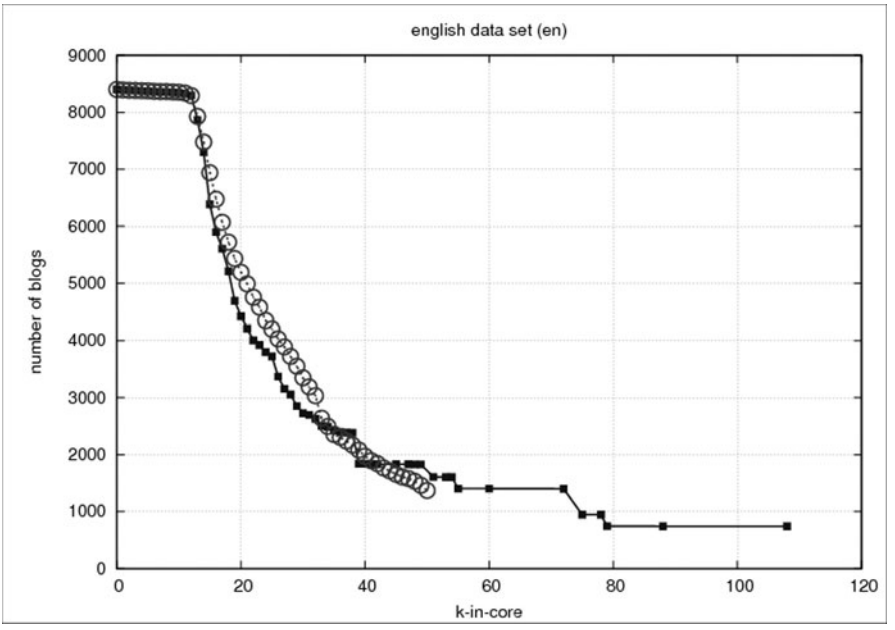


Fig. 5: In-cores for english blogs

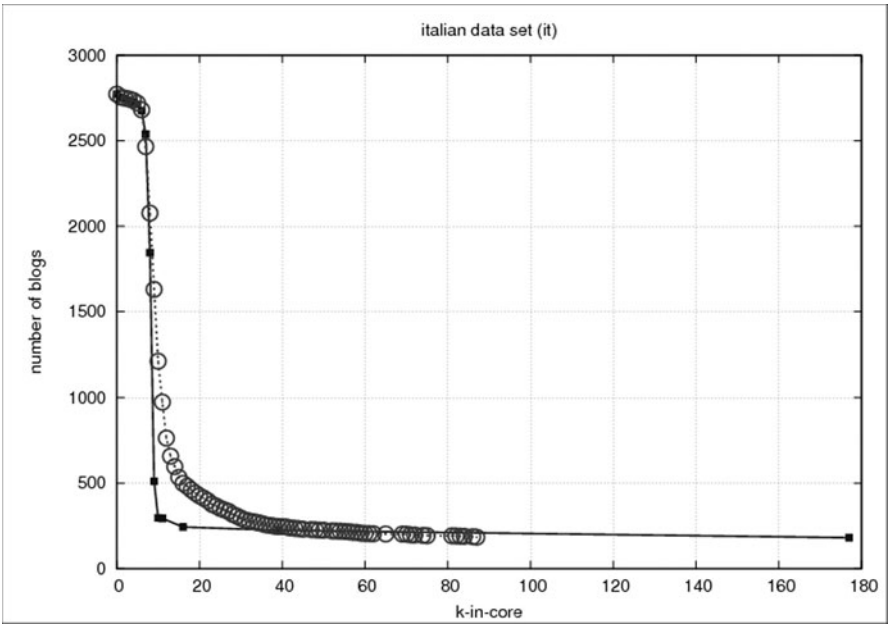


Fig. 6: In-cores for italian blogs

4.2 Comparing the Data Sets

In a second step, we compare the results of the different data sets with each other, and thus exploit the fact that, by language-specificness, we have similarly structured networks that can disguise anomalies in a network that do not occur the other ones.

When looking through the plots, one will immediately notice that the tendency towards this core-centralisation is different among the data sets. For the random networks, there is a correlation between the average degree and the curve of the expected k-in-cores. The lower the average degree, the steeper the curve falls, i.e. the less core-centralisation is normally expected, and thus, the resulting cores from the german blogs have to be judged differently than those of the english ones.

Furthermore, we notice that the german and the spanish blogs contain a very small core at their highest k, i.e. a 10-in-core of 25 german blogs and a 37-in-core of 39 spanish blogs, a phenomenon that does not appear in the other four data sets. A survey of the blogs in these small cores reveals two interesting explanations. The 25 german blogs all deal with cooking and recipes, and are well-interconnected. The 39 spanish blogs are all run by the

commercial blog network BlogsFarm[5], that runs about 50 blogs, and are thus nearly completely connected. The same explanation applies to the rest of the 78 blogs that form the spanish 28-in-core, these are run by the commercial blog network WeblogsSL[6], which maintains about 25 blogs. The arising question in both cases is, whether these blogs are only popular among themselves, due to commercial interests, or if they are also fulfilling the most important A-List characteristic, namely to be massively pointed to by other blogs from outside the core. This question cannot be answered by core-analsyis, but needs to be examined further, as we are going to outline in Section 5.

Another thing to notice is the much higher than expected maximum k in the spanish, the italian and the english data sets, which is very different from what is observed in the german, the portuguese and the french ones. This is an indicator for a large, well-interconnected group beyond the core-centralisation as emerged by the A-List phenomenon, according to our understanding. This issue has already been partially clarified for the spanish blogs, but in the english data set, we find 744 blogs that form a 108-in-core, and in the smaller italian data set we even find a 177-in-core of 181 blogs.

Despite the size of the english data set, this number appears too high for a sane community, and indeed we have found an interesting explanation. Our first suspicion, to have encountered a circle of spam blogs (splogs) did not hold, instead, this core is constructed by about 150 blogs that all include the "Blogging Chicks Blogroll"[7], a so-called *collaborative blogroll* with these 744 blogs, which aims to "take over the Internet, one blog at a time". This is a unique phenomenon in the english data set, which prohibits a reliable A-List detection with core-analysis only.

For the italian data set, the explanation is the same as for the spanish one, albeit on a significantly larger scale. The highest in-core is formed by blogs from the commercial blog network Blogosfere[8], which runs roughly 200 blogs on different topics. Here again, the same question of general popularity has to be examined.

We also notice a high dominance of one single blog-engine provider in the french data set, which is a unique phenomenon as well. From the 274 blogs in the french 17-in-core, 89% are hosted by *canalblog.com*, opposed to 68% in the whole data set of 3402 blogs. A survey revealed no signs for a systematic favorisation between these blogs, so we regard it as a purely cultural phenomenon and consider the french blog data set to be free of anomalies. The same holds true for the portuguese data set, which also seems to be free of anomalies beyond the expected core centralisation phenomenon. Consequently, these two data sets will serve as references for a sane manifestation of the core centralisation phenomenon for A-List detection.

[5] http://blogsfarm.com/about/

[6] http://www.weblogssl.com/quienes-somos

[7] http://bloggingchicks.blogspot.com/

[8] http://blogosfere.it/about.html

4.3 Comparison with the Core/Periphery Model

In a third step, we validate the approach by comparing it with Borgatti and Everett's core/periphery model. In the case of a directed network, the variation of their "asymmetric model" is the one relevant to us. Their example, citations among 20 scientific journals, is comparable in the sense of a core/periphery structure, and also emerges something similar to an A-List, namely a subset of journals that fulfills the three A-List characteristics reasonably well.

With our in-core analysis, we detect a 4-in-core that contains 6 journals. This is one more than identified by them as "the core" (see [4], p. 385). The journal in question, "ASW", is included in our 4-in-core, because it is referenced by four other journals from that core. On the other side, it is not included by the core/periphery model, because there are no links at all from the periphery to that journal, and thus, it cannot be considered as an authoritative one with confidence. This limitation of our core-analysis towards anomalies against the first A-List characteristic has already been observed in the blog data sets, and is independently confirmed here. As mentioned before, this problem is addressed in the following section.

5 Identifying A-List Blogs

5.1 Constraints

In order to reliably detect A-List blogs, all three characteristics must be fulfilled. The core-analysis can provide valuable insight to the second characteristic. However, the first and the third characteristic require an analysis of the core's relation to the periphery, which is not directly addressed by our method. In fact, the emerging cores do comply to all three characteristics in the random networks, but not necessarily in the real-world networks with their special anomalies, as we could see in the previous section.

The highest k-in-core of the french data set, a 17-in-core with 274 nodes, and the highest k-in-core from the portuguese data set, a 20-in-core with 209 nodes, are the only ones that are free of such anomalies and can be immediately used as an A-List represenation. For all other original data sets, a combination with further analyses is required, where different methods have to be considered and compared.

In a first step towards this goal, we try to explicitly quantify the anomalies observed in the four data sets by measuring how well core members comply to the expected characteristics of core centralisation as observed in the french, the portuguese and the random networks.

We have to be aware of the fact that the long tail is missing in our data sets, due to the nature of the data acquisition method. For example, the number of incoming links from the collaborative blogroll in the english data set is higher than any number of incoming links a blog receives from the periphery. This would not remain true in a larger data set with many more blogs in the lower cores. In order to detect the anomalies properly, we thus have to find a metric that is immune to the absence of periphery blogs under the assumption that these blogs are connected to the core as expected.

5.2 Structural Analysis

Members of higher k-in-cores in average receive more incoming links from the rest of the network than members of lower k-in cores do, which conforms to the first A-List characteristic. This is true for all random networks, but in the original blog data sets, this is true only for the french and the portuguese ones. When not true. it is an indicator for the fact that the higher cohesion is only added by a local effect, as the recipe and cooking theme in the german 10-in-core for example (see Section 4.2).

This would work for the german and the spanish data sets, but the average number of incoming links is not immune to the missing long tail links, as the nodes in the highest in-cores of the english and the italian data sets have the highest average in-degrees, despite being referenced less often from the long tail than many nodes in lower in-cores. This is a result of their extremely high linking amongst each other. To eliminate this effect of intra-core links, we can count only incoming links from outside the node's k-in-core. This in turn does not account for the iterative nature of nested cores. With this metric, we still see nodes with little incoming links from the periphery, but with high in-degrees from outside their k-in-core, because a large portion of their cohesive subgroup forms an in-core with a slightly lower k, e.g. a $(k-1)$-in-core.

5.3 Core Independency

Our final solution is to weight each incoming link of a target node based on the core-distances between the target node and the source nodes, i.e. the lower the in-core of the source node relative to the in-core of the target node, the more valuable that link is for determining the effect of the first A-List characteristic.

We call this metric *core independency*, as it measures how little a node's authority depends on its fellow core members.

Given a function $k(v)$ returning the maximum k for which a node v is a member of a k-in-core, we can define the core independency $indep(v)$ of a node v with $k(v) \geq 1$ as follows.

$$indep(v) = \sum_{i=0}^{k(v)-1} \frac{k(v) - i}{k(v)} \cdot \frac{|\{(s,t) \in E \mid t = v \wedge k(s) = i\}|}{indeg(v)} \tag{1}$$

For nodes that are not members of any core, the independency is 0 by definition. The values of this metric will be in the interval $[0, 1[$, and the complementary metric *core dependency* can be defined as $dep(v) = 1 - indep(v)$.

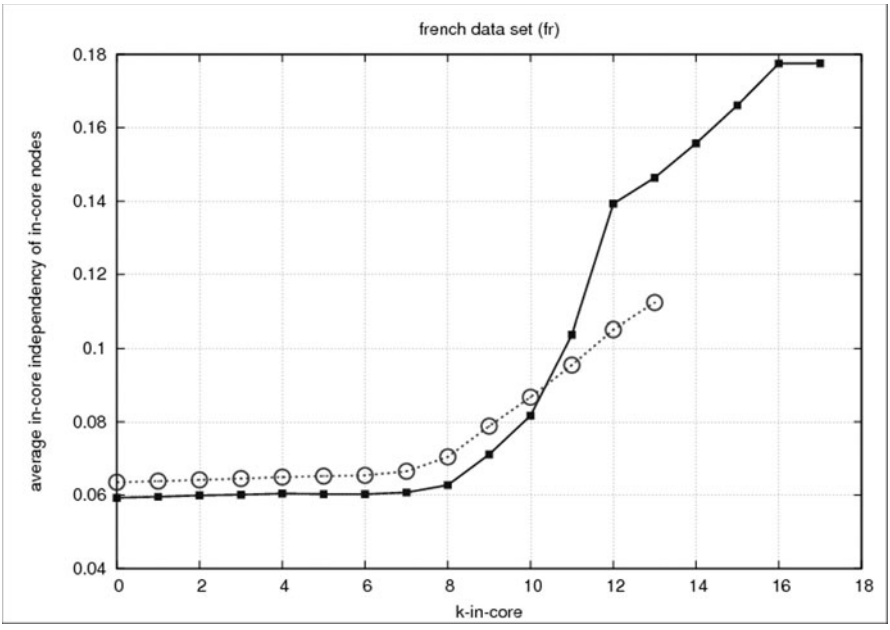


Fig. 7: Average independencies in french in-cores

Figures 7 to 12 plot the core independency metric for all of our data sets, whereby the x-axis denotes the k-in-core and the y-axis denotes the corresponding average core independency of the core members. Again, the circles represent the results from the random network and the squares represent the values of the original data sets.

Apparently, this metric is capable to visualise all the different anomalies we observed in Section 4.2. If more periphery blogs were present, the independency values in higher k-in-cores would increase in average, but the curve shapes would remain the same.

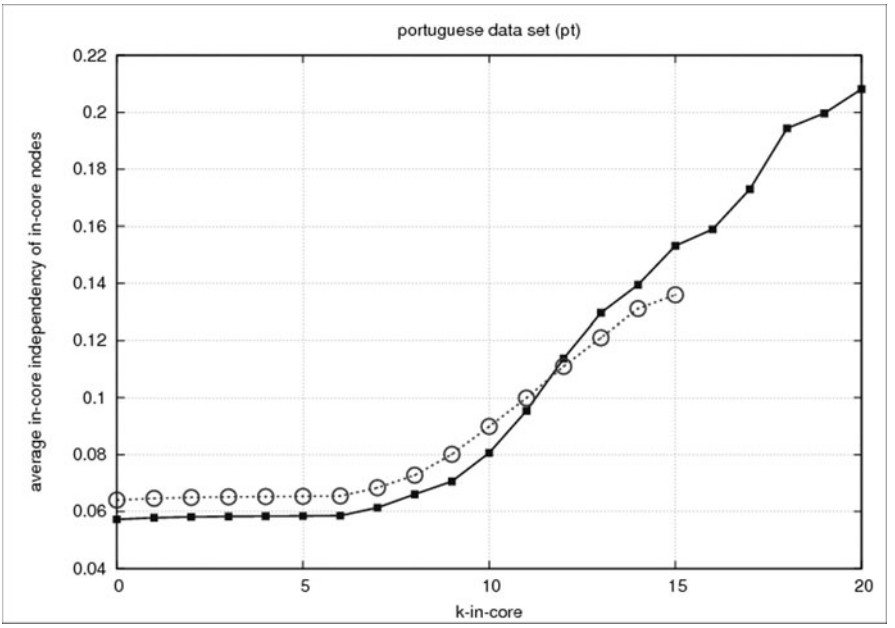


Fig. 8: Average independencies in portuguese in-cores

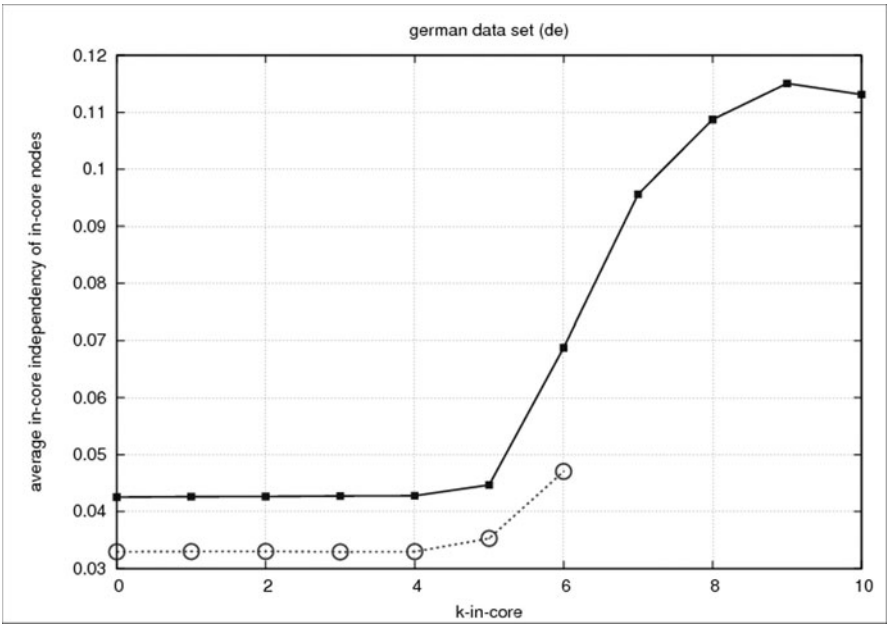


Fig. 9: Average independencies in german in-cores

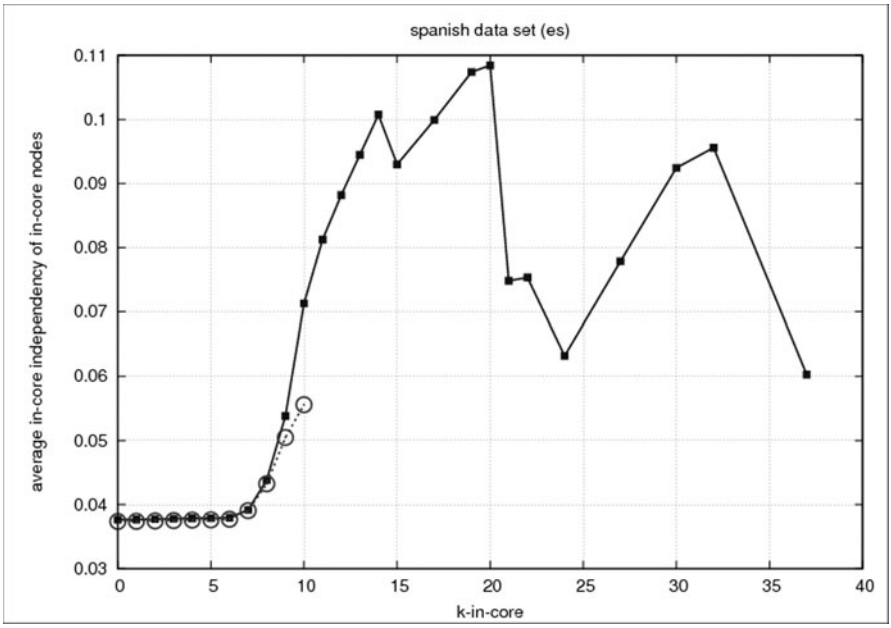


Fig. 10: Average independencies in spanish in-cores

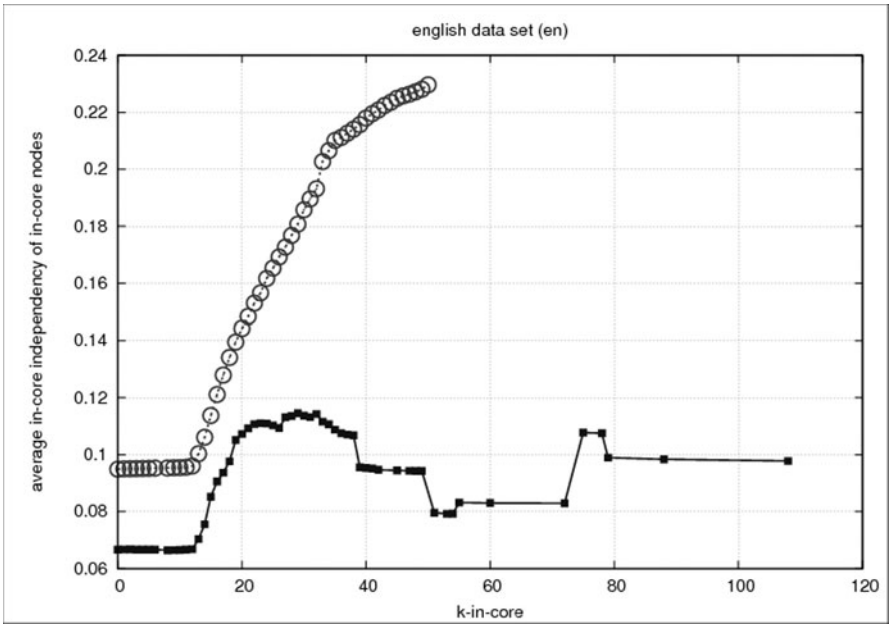


Fig. 11: Average independencies in english in-cores

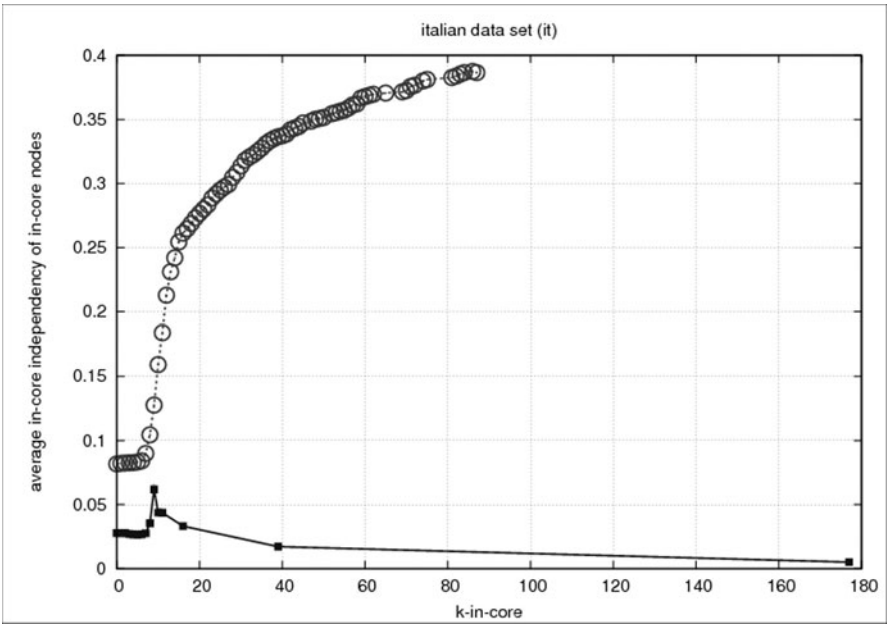


Fig. 12: Average independencies in italian in-cores

As a metric for individual nodes, the core independency can be used to remove nodes under a certain threshold from the final A-List candidate list, or "the core" according to the interpretation of Borgatti et al [4]. In fact, the problematic journal "ASW" mentioned in Section 4.3 has a core independency of 0, which makes it a candidate for removal no matter what threshold above 0 will be chosen.

6 Conclusion

In this chapter, we have developed and applied an efficient variation of core-analysis to blog data-sets of different languages. The analyses revealed a general tendency towards more-than-average core-centralisation, as well as a number of interesting phenomenons, that were not all publicly known by now. In this process, the different language data-sets proved to be highly useful for cross-checking. We also have shown that core-analysis, in conjunction with the newly proposed metric of *core independency*, can be used to detect A-List blogs in a more sophisticated way than it is done up to date. This is a promising step for improving rankings in the future.

The results also bring up new questions in directions other than A-List detection. The individual phenomenons of the different languages, e.g. the local dominance of a certain blog-engine provider or the nature of collaborative blogrolls, could be related to cultural specialities and allow conclusions about different blogging cultures.

References

1. A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
2. R. Blood. *The Weblog Handbook: Practical Advice on Creating and Maintaining Your Blog*. Perseus Books, 2002.
3. B. Bollobas. *Random Graphs*. London: Academic Press, 1985.
4. S. P. Borgatti and M. G. Everett. Models of core/periphery structures. *Social Networks*, 21:375–395, 1999.
5. A. Delwiche. Agenda-setting, opinion leadership, and the world of web logs. *First Monday*, 10(12), 2005.
6. P. Doreian and K. L. Woodard. Fixed list versus snowball selection of social networks. *Social Science Research*, 21(2):216 – 233, 1992.
7. P. Doreian and K. L. Woodard. Defining and locating cores and boundaries of social networks. *Social Networks*, 16(4):267 – 293, 1994.
8. S. C. Herring, I. Kouper, J. C. Paolillo, L. A. Scheidt, M. Tyworth, P. Welsch, E. Wright, and N. Yu. Conversations in the blogosphere: An analysis "from the bottom up". In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, page 107.2. IEEE Computer Society, 2005.
9. C. Marlow. Audience, structure and authority in the weblog community. In *Proceedings of the International Communication Association Conference*, 2004.
10. D. Obradovic and S. Baumann. Identifying and analysing germany's top blogs. In *Proceedings of the 31st German Conference on AI*, pages 111–118. Springer, 2008.
11. D. Park. From many, a few: Intellectual authority and strategic positioning in the coverage of, and self-descriptions of, the "big four" weblogs. In *Proceedings of the International Communication Association Conference*, 2004.
12. S. B. Seidman. Network structure and minimum degree. *Social Networks*, 5:269–287, 1983.
13. C. Shirky. Power laws, weblogs, and inequality, February 2003.
14. F. Tricas, V. Ruiz, and J. J. Merelo. Do we live in an small world? measuring the spanish–speaking blogosphere. In *Proceedings of the BlogTalk Conference*, 2003.
15. D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, (393):440–442, 1998.

Social Physics of the Blogosphere

Capturing, Analyzing and Presenting Interdependencies within a Single Framework

Justus Bross, Keven Richly, Patrick Schilf, and Christoph Meinel

Abstract It was already shown on several occasions that it can be highly meaningful for individuals, institutions or even governments to find ways and measures in order to extract reliable and insightful trends, opinions or particular pieces of information out of the blogosphere. However, it is increasingly difficult if not impossible for the average internet user and sympathizer of weblogs to grasp the blogosphere's complexity as a whole, due to thousands of new weblogs and an almost uncountable number of new posts adding up to the before-mentioned collective on a daily basis. Mining, analyzing, modeling and presenting this vast pool of knowledge in one central framework to extract, exploit and represent meaningful knowledge for the common blog user forms the basis of this paper. The result of the corresponding long-term research initiative presented here is BLOGINTELLIGENCE. It is an integrated blog analysis framework with the objective to leverage content- and context-related structures and dynamics residing in the blogosphere and to make these findings available in an appropriate format to anyone interested. We hereafter refer to these structures and dynamics as social physics of the blogosphere.

Justus Bross,
e-mail: justus.bross@hpi.uni-potsdam.de

Keven Richly,
e-mail: keven.richly@student.hpi.uni-potsdam.de

Patrick Schilf,
e-mail: patrick.schilf@student.hpi.uni-potsdam.de

Christoph Meinel,
e-mail: office-meinel@hpi-web.de

Hasso-Plattner-Institut, Internet Technologies and Systems,
Prof.-Dr.-Helmert-Strasse 2–3, 14482 Potsdam, Germany

1 Introduction

Since the end of the 90's weblogs - commonly known as blogs - have evolved to become an essential component of today's cyber culture [1]. In the year 2008, the worldwide number of blogs has increased to a total in excess of 133 million. Compared to around 60 million blogs in the year 2006, this constitutes their increasing importance in today's internet society on a global scale [2,3]. Technically speaking, weblogs are an easy-to-use but highly configurable web-enabled Content Management System (CMS), in which dated articles - also known as posts - and comments on these posts, are presented in reverse chronological sequence [4]. With popular blogging software systems such as Wordpress or Blogger.com that are intuitively to handle, any internet user with less than average technical basic knowledge is nowadays able to write anything at any time, about whom and about whatever he likes to. Basically everyone is therefore able to reach anybody with his personal opinion that has an internet connection and speaks the same language. A weblog's point of origin is meanwhile indefinable since their potential areas of application are numerous - beginning with personal diaries, reaching over to knowledge and activity management platforms, and finally to enabling content- related and journalistic web offerings [5]. A continuative differentiation of this media-phenomena seems however inevitable to cope with the overall complexity of the blogosphere. This is among others due to an elevated public awareness, diverse topics covered and the different fields of interest blogs are employed in. Also their global interconnectedness and the corresponding aggregation of individual knowledge create a gigantic and constantly changing archive of open source intelligence [6]. This is why weblogs are often grouped into the family of "social software" [7].

1.1 The Bigger Picture

Single weblogs are embedded into a complex superstructure: an independent and segmented public that dynamically evolves and functions according to its own rules and with ever-changing protagonists, an exceptionally interlinked network also known as the "blogosphere". A single weblog is embedded into this network through its track- or ping backs, the usage of hyperlinks in terms of referrers and common links to other blog instances as well as its so-called "blogroll" - a blogosphere-internal referencing system [8]. Internet-specific connectivity mechanisms like permalinks and feeds support this interconnected super structure.

The past has shown on several occasions that it might prove highly meaningful for a multitude of individuals, institutions or even governments to find ways and measures in order to extract reliable and insightful trends, opinions or particular pieces of information out of the blogosphere. No matter for what

purpose the vast pool of information within the blogosphere was ultimately retrieved: through their direct, informal and unadorned mode of operation, weblogs could often serve as the fastest provider of insight information about technical product innovations, politics or coverage on a multitude of other topics.

However, it proves to be increasingly difficult for the average internet user and sympathizer of blogs to grasp the blogosphere's complexity as a whole, yet due to the fact that thousands of new blogs and an almost uncountable number of new posts add up to the before-mentioned collective on a daily basis. Up-to-date and regular content generation has become a form of credo for existing weblogs to survive in the blogosphere.

1.2 Research Rationale

The absence of any centralized control instance, which is usually regarded as the blogosphere's biggest congeniality, is its major shortcoming in the be-fore mentioned context: Mining, analyzing and modeling this vast pool of unstructured information in one central framework simply seemed virtually impossible so far. Extracting meaningful knowledge out of this pool to leverage content- and context-related structures residing in the blogosphere, thus forms the ultimate objective of the long-term research project presented in this paper.

Intuitively, one could apply some of the numerous existing data or web mining approaches, blog analysis methods and visualization techniques that meet our be-fore-mentioned general requirements of a central blog intelligence framework. Most of these numerous existing research project's are indeed meaningful in their respective fields and with their particular focus, but unfortunately do not meet our requirements to a full extend. We therefore rearranged all existing research projects with our own findings to have a perfectly concerted research framework from the projects beginning. All these research efforts stand for a new trend that some call "Computational Social Science", and others "Social Physics" or "Network Science". Their common approach is founded in the conviction that modern society is a network of social atoms. Social phenomena, such as the formation of millions of weblogs to a highly interconnected blogosphere, should according to this conviction be explainable with methods of natural science [9]. These researchers believe, that interconnections (e.g. of the blogosphere) are a result of social chain reactions that are based upon simple principles. It is the ambition of this research initiative to ultimately shed some light on the social physics of the blogosphere.

1.3 Research Overview and Chapter Arrangement

The subsequent section of this paper starts with introducing the most notable existing blog analysis engines already on the market that can to some extent be compared to an integrated blog intelligence service as outlined within our research project. Section three then outlines this research framework in detail. It provides a conceptual overview of our three-years-long research initiative by introducing its three major research phases "Extraction", "Analysis" and "Visualization". These three phases are elaborated on in more detail in the subsequent sections four, five and six. Section four about the "Extraction"-phase summarizes our research findings gained through the preceding corresponding research as well as the implementation of a crawler that is capable of collecting all the data necessary for the research discussed in the following. The section called "Analysis" discusses all those components and metrics that we consider to be crucial for a blog analysis framework as we plan it. It elaborates upon the shortcomings of these external projects but discusses at the same time meaningful single aspects that are worthwhile to be considered for further research. These insights are then merged with our own findings from three years of intensive research activities in the field of the blogosphere to ultimately provide a solution to our overall research objective in sections six. Here, we discuss ways and measures of how the information collected and all the single metrics analyzed during the two preceding stages can be most meaningfully presented and visualized in only one central instance to potential users of our blog intelligence service. The conclusion in section seven provides the final comparison of our framework with existing services. The upcoming project milestones as well as recommendations for our further research are equally given in this section, followed by the reference list section eight.

2 Related Work

There are several commercial as well as non-commercial institutions offering an integrated blog analysis engine that is - according to some characteristics - similar to our own research project. We compared these services on the basis of all those features and metrics that we consider essential for such a service. Even though figure one gives a thorough overview to what extent currently existing engines provide these basic features, we would briefly like to mention the most important features of these engines in the following.

One of the most comprehensive services is BLOGPULSE. Using this Internet platform, users can retrieve all kinds of different analyses about single weblogs and the blogosphere as a whole. Next to a keyword-based search for single posts, BLOGPULSE also offers trend-analysis in the blogosphere. The frequency of an arbitrary term can hereby be displayed for a predefined

period of time and ultimately compared with others. BLOGPULSE also offers up-to-date rankings of weblogs. Their specialty in this regard is to offer charts about various popular metrics such as "Prominently featured people in today's blog buzz", "Most-cited posts among today's bloggers" or "Top 40 blogs, based on links by other bloggers". These charts are generated daily and can also be displayed for periods in the past. A conversation tracker is also included within BLOGPULSE. In this regard, conversations are sequences of posts in different weblogs that are about a similar topic and link to each other. BLOGPULSE also builds up profiles about single weblogs and can consequently provide basic information over time such as the individual ranking of a weblog, the number of citations or posts and comments published. Another interesting element is the "Neighborhood Feature". Here, the user is given notice of about ten other weblogs that cite the same sources and publish posts about similar topics.

With its "authority value", TECHNORATI offers a similar ranking of weblogs within the blogosphere as BLOGPULSE. It is furthermore a specialty of TECHNORATI to provide a post- and weblog-specific search functionality. Similar to the ICEROCKET service, TECHNORATI offers a commercial interface that allows users to run own analyses over their data. Other features of ICEROCKET are their own ranking score and a trend analysis within the blogosphere, as well as basic information about single weblogs.

SPINN3R differs substantially from all other services investigated, since it is a service in cooperation with the universities of Stanford, Harvard and Carnegie Mellon that only provides access to its data archive with costs. Another interesting service is POSTRANK. It calculates its ranking metric on the basis of user interaction with content of a specific blog, and not - as all other services - on the basis of incoming links. User's reaction on this content in different social platforms such as Twitter is also monitored in this regard. However, its search functionality only allows for the search of posts. BLOGSCOPE is a service developed by the University of Toronto. Next to the provision of its own ranking algorithm and a trend analysis, it allows users to filter their search results with individually chosen ranking metrics. AMATOMU is a service that monitors the South African blogosphere. In contrast to most of the other engines, it offers a search functionality that allows for the search of posts, blogs and even individual bloggers. Next to a trend analysis, it offers a ranking algorithm that is determined out of the three metrics reads, posts and links. Common blog analysis is also realized by TWINGLY, a service that provides basic information about single weblogs as well as its own ranking metric. Since the German blogosphere will form the research basis for our own blogosphere analysis engine BLOGINTELLIGENCE, we should also mention "Deutsche Blogcharts (DBC)[1] " at this point. It provides an often cited ranking chart of (and for) the German-speaking blogosphere. Their ranking metric is however based on the data of ICEROCKET, which

[1] http://www.deutscheblogcharts.de

is one the Blogosphere analysis engines discussed before. Interestingly, DBC used to employ the data of TECHNORATI as the basis for its ranking charts since early 2006, but ultimately decided to switch to ICEROCKET due to fundamental reasons of intransparency what TECHNORATI's methods regarding data collection, storage or analysis is concerned [10]. Another service that provides no information at all in this regard is GOOGLEBLOGS. On top of utmost transparency, BLOGINTELLIGENCE intends to provide all those web-log-data analyses within one single framework that no other integrated engine can provide in the aggregate at this moment. Refer to Figure 1 for an overview of the data analyses and read sections 5.1 and 5.2 to get the details of the underlying metrics employed in each analysis.

3 Framework

The analysis of data generated within and from the blogosphere's network can be insightful for numerous reasons and for a high diversity of interest groups. The blogosphere however represents a partition of the WWW that dynamically evolves and functions according to its own rules. These different characteristics are the foremost reason why existing mining and analyses methods could not be equally applied within the blogosphere and the WWW [11]. We therefore develop our own framework that comprises the three main components "Extraction", "Analysis" and "Visualization" (see figure 2).

The Extraction phase is about getting the necessary information for future analyses out of the blogosphere's cyberspace. We make use of a crawling framework to get the information and store it in a database in proper order. Since traditional crawler implementations do not fully consider the particularities of weblogs as opposed to traditional websites, we had to implement a crawler purpose-built for the blogosphere on our own. The subsequent section will elaborate further upon the before-mentioned blogosphere's particularities as well as on the crawler's implementation details. The second (and central) part of our framework - the "Analysis" - is concurrently performed while the crawler continuously collects new information. Data analyzers are working on the information stored in the database and process that information for the third part of the framework - the Visualization part. Due to the modular built-up of the data analyzers, it is at all times possible to add new or modified data analyzers to the system, or delete those that are not of interest anymore. The data analyzers can generally be divided in two main categories. Network analyzers investigate the linking structures within the blogosphere and can, for instance, provide crucial information about relationships of different weblogs or communities of interest. These dependencies are typically investigated by means of graph analyses. Content analyzers in turn make use of common text-mining techniques to allow for the content-related analysis of weblogs. On the basis of these analyses, we can for instance make state-

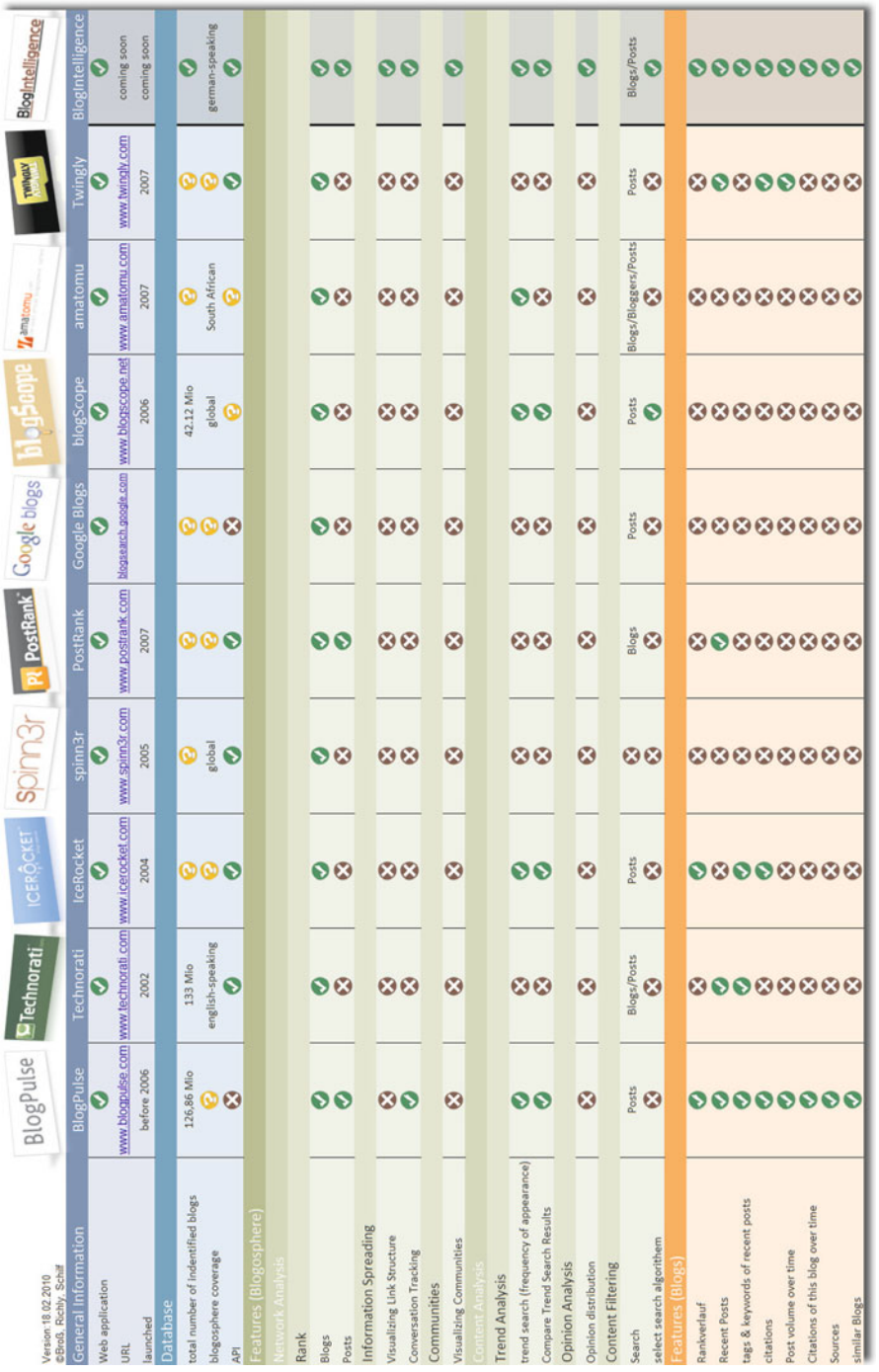


Fig. 1: Comparison of Blogosphere Analysis Engines (own visualization)

ments about content-related correlation of different weblogs, or the number of bloggers writing about similar topics.

However, the information generated by the data analyzers still not provides the user the opportunity to identify relevant blogs, important communities, key bloggers in a partial blogosphere or topics of individual interest. The final step of our framework - the "Visualization" - thus provides the interface between the processed information and the user. It allows users to browse the pre-processed information of the data-analyzers in an unlimited, personalized and intuitive way. All metrics of BLOGINTELLIGENCE are allocated to the user in one central web-enabled interface. For reasons of manageability and due to the enormous amount of information that has to be made available, we subdivided the visualization interface into three layers. The top abstraction layer visualizes the general interdependencies and linkages of weblogs in the blogosphere. All the information regarding single weblogs and their content is visualized in the layer underneath.

4 Extraction: Data Elements and Crawler Implementation

Blogs consist of diverse, and more or less clearly identifiable elements that con-tain the information that is of interest for us and which consequently need to be completely collected by our crawler [12].

4.1 Information Elements

Even though already shortly introduced in the first section, these specific elements of interest are summarized in the following before we elaborate upon the implementation details of the blog-crawler as outlined by Bross et al. [8].

Posts. A post - the central element of a weblog - is an article that is annotated with a timestamp, suitable tags and classified into appropriate categories. On top of textual elements, also multimedia data can be attached to a post. When published, posts are displayed in reverse chronological order on the starting page, meaning that the most recent post is always presented at the top of the page.

Comments. Comments are given on postings of other bloggers. They thus represent the most basic form of interaction in weblogs. Several consecutive comments on one post are called "threads". Comments are not necessarily instantly published, but firstly moderated by the weblog administrator to filter out spam-comments.

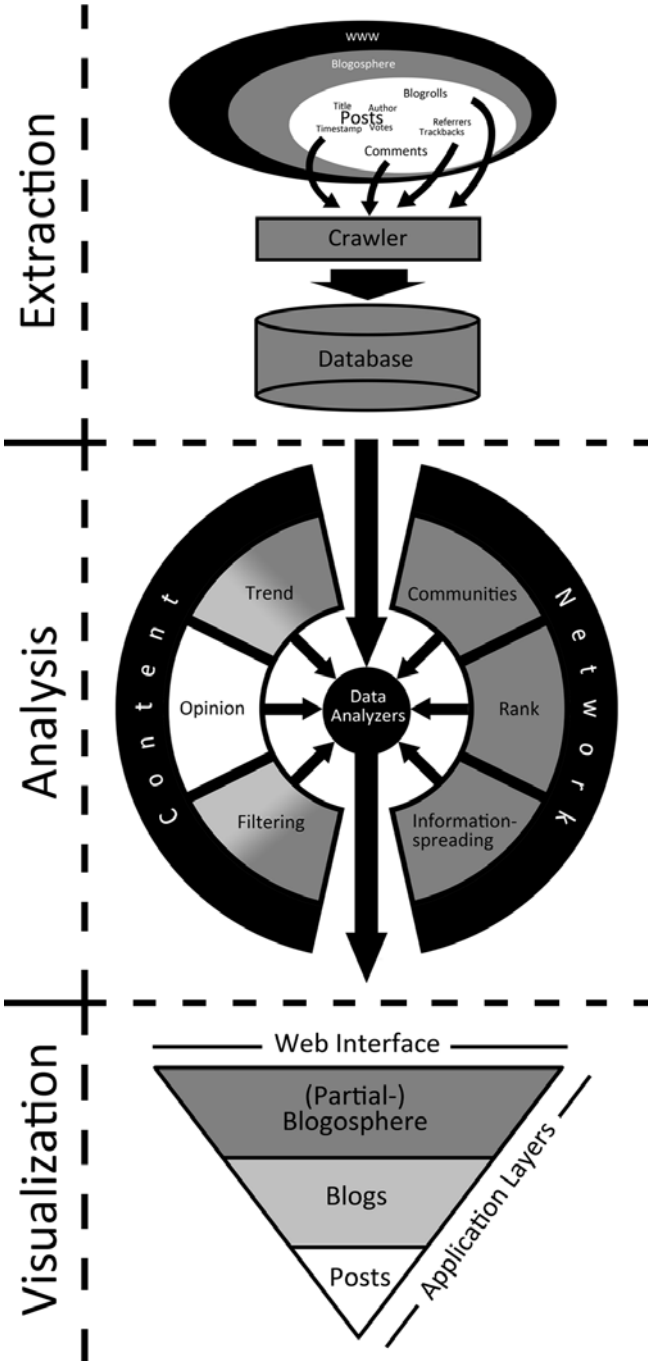


Fig. 2: Research Overview and Concept (own visualization)

Blogroll. Almost every weblog has a listing of other weblogs that the author regularly reads. It usually covers similar fields of interests and is placed in the sidebar. This form of interlinking evolved in the early days of the medium both as a navigational tool for readers to find blog-platforms with similar topics as well as some sort of social acknowledgement within the community. In some hosted services, the blogroll is a core part of the interaction, allowing users to be notified when their friends pulish a post or even create a group dialog represented by the sum of the group's individual blogs (also refer to track- or pingbacks).

Permalinks. It is an essential for the interaction within blogs to allow others to reference specific post or comments in other weblogs. These entry reference points require a permanent and explicitly identifiable web-address. This element represents the core of almost all weblog systems today.

Track- or Pingbacks. Trackbacks allow web authors to request notification when somebody links to one of their postings. This enables authors to keep track of who is linking, and so referring, to their articles These explicit referrers are displayed similar to comments attached to a post. Pingbacks are a special form of the trackback-principle that allows for an automatic detection of trackbacks which do not need to be typed in manually anymore. These automated pingbacks are only supported by some weblog software programs, such as Serendipity, Wordpress, CuteNewsRU, Movable Type, Typo, Telligent among others.

Feeds. Blog content, such as posts or comments, can be distributed in a standardized format such as RSS or ATOM throughout the web via so-called Feeds. In the blogosphere, RSS-feeds are usually provided whenever a new post or comment is published in weblogs. Due to the standardized format of RSS-feeds, machines or program routines can automatically analyze them, and are consequently able to provide subscribers with updated and current content of these feeds. You could thus say that the sum of all feeds represents the network's entire structure, because of which feeds are most important information elements for our crawler.

Metadata. On top of the data elements discussed before, additional information about the underlying blog system is collected as well.

4.2 Crawler Action-Sequence

The crawler starts his assignment with a predefined and arbitrary list of blog-URLs. Usually this predefined catalog is the top 100 list of a blog search engine like Technorati. It downloads all available post- and comment-feeds of these blogs and stores them in a database. It than scans the feed's content for links to other resources in the web, which are then also crawled and equally downloaded in case these links point to another blog. Again, the crawler starts scanning the content of the additional blog feed for links to

additional weblogs. The crawler repeats this iterative process till it comes up against a link that was either already scanned, or till he comes across so-called "isles" - a smaller network of blogs that only link to each other and have no connection to the rest of the blogosphere. To avoid that the crawler gets stuck on one of these isles, we include blog URLs from different geographical regions, as well as blogs that cover diverse and unrelated topics as regards content in the arbitrary starting list. This furthermore increases the odds that the crawler covers the whole of the blogosphere within a minimum of time. With maximal content related diversity in the arbitrary starting list, data can also be meaningfully analyzed in an early stage. By following-up links, it can be guaranteed that any URL of the worldwide top weblogs will sooner or later be crawled. The representation of the most influential opinion leaders is therefore feasible.

4.3 Recognizing Weblogs

Whenever a link is analyzed, the crawler first of all needs to assess whether it is a link that points to a weblog, and also with which software the blog is created. Usually this information can be obtained via attributes in the metadata of a weblogs header. It can however not be guaranteed that every blog provides this vital information for us as described before. There is a multitude of archetypes across the whole HTML page of a blog that can ultimately be used to identify a certain class of weblog software. By classifying different blog-archetypes beforehand on the basis of predefined patterns, the crawler is then able to identify at which locations of a webpage the required identification patterns can be obtained and how this information needs to be processed in the following. The crawler knows how to process the identification patterns of the most prevalent weblog systems around: Wordpress[2] , MovableType[3] , Blogger.com[4] , Serendipity[5] among others [13]. In the course of the project, Identification patterns of other blog systems will follow. In a nutshell, the crawler is able to identify any blog software, whose identification patterns were provided beforehand.

[2] https://wordpress.org

[3] http://www.movabletype.org

[4] http://blogger.com

[5] http://www.s9y.org/

4.4 Recognizing Feeds

The recognition of feeds can similarly to any other recognition-mechanism be configured individually for any blog-software there is. Usually, a web service provider that likes to offer his content information in form of feeds, provides an alternative view in the header of pages, defined with a link tag. This link tag carries an attribute (rel) specifying the role of the link (usually "alternate", i.e. an alternate view of the page). Additionally, the link tag contains attributes specifying the location of the alternate view and its content type. The feed crawler checks the whole HTML page for exactly that type of information. In doing so, the diversity of feed-formats employed in the web is a particular challenge for our crawler, since on top of the current RSS 2.0 version, RSS 0.9, RSS 1.0 and ATOM among others are also still used by some web service providers. Some weblogs above all code lots of additional information into the standard feed. Momentarily, our crawler only supports standard and well-formed RSS 2.0 formats, of which all the information of our currently employed object-model is readout. It is the aim of our project team to include as many RSS-formats as possible in the future.

4.5 Storing Crawled Data

Whenever the crawler identifies an adequate (valid) RSS-feed, it downloads the entire corresponding data set. The content of a feed incorporates all the information necessary, to give a meaningful summary of a post or comment - thus a whole weblog and ultimately the entire blogosphere. General information like title, description, categories as well as the timestamp indicating when the crawler accessed a certain resource, is downloaded first. Single items inside the feed represent diverse posts of a weblog. These items are also downloaded and stored in our database using object-relational mapping[6] (refer to figure two in the appendix). The corresponding attributes are unambiguously defined by the standardized feed formats and by the patterns that define a certain blog-software. On top of the general information of a post, a link to the corresponding HTML representation is downloaded and stored as well. In case this information is not provided in the feed information of a blog provider, we are thus still able to use this link information at a later point for extended analyses that would otherwise not be possible. Comments are the most important form of content in blogs next to posts, and they are usually provided in form of feeds as well. However, we do need to take into account that a comment's feed-information is not always provided in the same form by all blog software systems. This again explains why we use pre-defined distinct blog-software classes in order to provide the crawler with the necessary

[6] https://www.hibernate.org/

identification patterns of a blog system. Comments can either be found in the HTML header representation or in an additional XML attribute within a post feed. Comment feeds are also not provided by every blogging system. With the predefined identification patterns, our crawler has however a time issue, since it has to download the essential information of the comment and store it in our database. Another important issue is the handling of links that are usually provided within posts and comments of weblogs. In order to identify network characteristics and interrelations of blogs within the whole of the blogosphere, it is not only essential to store this information in the database, but to save the information in which post or comment this link was embedded.

4.6 Refreshing Period of Crawled Data

How often a single blog is scanned by our crawler should depend on its cross-linkages with other blogs. Blogs that are referenced by other blogs via trackbacks, links, pingbacks or referrers are thus visited with a higher priority than others by the crawler. Well-known blogs that are referenced often within the blogosphere are also revisited and consequently updated more often with our original algorithm. It can be considered possible that with this algorithm blogs of minor importance are visited rarely - a side-effect that we do not consider to be limiting at this time. Since the blogosphere is constantly changing with new blogs being setup and other blogs disappearing (as can be inferred from the changed starting list), it is of crucial importance that the crawler preferable also finds new blogs and not only refreshes existing ones. We realized this requirement on the basis of "priorities" - hereafter referred to as "Prio". A Prio is the number of hops necessary to get from the initial URL starting page to a particular blog, whereas all blogs within the starting list do have a Prio-value of 0. All those links that are collected on the front pages of one of the starting list blogs thus have a Prio-value of 1. To guarantee that the crawler neither only updates those blogs it already found, nor merely tries to find new blogs without updating the information of the existing ones, new jobs to be crawled are scheduled as follows: There are several parallel working analyzers and a scheduler that determines which job will be processed next by the analyzers. The scheduler processes all job with Prio = 0 on a daily basis. After that, all those links with Prio=1 that point to other blogs are also processed on a daily basis. At the time the analysis of blogs with Prio=1 is completed, the scheduler assigns two thirds of those analyzers available to analyze blogs with Prio = 2 that were not analyzed for more than a week. The remaining third of analyzers is assigned with new jobs. At the time these jobs are completed as well, one third of those analyzers available are assigned with jobs that point to blogs of Prio > 3 that were not processed for more than a week. The remaining analyzers are than equally filled up with new

jobs. When all blogs in the database are updated, the scheduler assigns all analyzers with new jobs that were not visited so far.

5 Analysis: Integration of Proprietary and Existing Research Efforts

This section focuses on the discussion about how the data collected by the crawler (refer to section four) can and should be analyzed to extract, exploit and represent meaningful knowledge and to leverage content and context-related structures and dynamics of partial blogospheres. There can be a multitude of data analyzers that could process the information as collected by the crawler (refer to "Extraction" in figure two). Our research supports the overall subdivision of this multitude in generally two main categories, namely "content analyses" and "network analyses" (refer to "Analysis" in figure two). Content Analysis relies on rather traditional text mining techniques like Vector Machines, Feedforward/Backpropagation neutral networks or the Naive Bayesian method to provide insights such as content-related accordance between different blogs or the number of bloggers that write about a similar topic - only to name a few. Network Analysis in turn focuses on dependencies and interlinkages between weblogs and is consequently rather interested in insights about the overall network of the blogosphere and not in single weblogs to discover - for instance - the most influential blog-platforms in the blogosphere or make a statement about social distances.

5.1 Network Analysis

To gain general insights about interconnection within the blogosphere, you will need a basic structural analysis tool based upon a graph-based interpretation that visualizes dependencies by making use of all types of linking information as described in section 4.1. A project that took this approach was VISUAL NEIGHBORHOOD, a tool that visualizes link structures of single weblogs by making use of spider web graphics [14][15][16]. Link structure was equally analyzed by Herring and his colleagues to examine dependencies between weblogs [17]. Their study empirically investigates the extent to which (and in what patterns) blogs are interconnected. Their visualization of link patterns in conjunction with quantitative social network analysis as well as qualitative analysis of references and comments found out that the most influential blogs, commonly known as "A-list" blogs, are overrepresented in the network although other groupings of blogs are more densely interconnected.

Information Diffusion Analysis

Several scholars undertook a form of information diffusion analysis on the basis of the before-mentioned structural analysis, to find out in what way, to what extend and how fast information spreads from its original source to other blog instances. A similar approach, which became known as the "infection analysis" and originally used to analyze the diffusion of epidemics, was successfully applied to the blogosphere to track the linking between single blog posts with the help of so-called "infection trees" [18, 16, 19, 20]. By observing these trees you could assess in what way information diffused throughout the blogosphere and see which blog "infected" another one in passing on the information of the original blog post [19, 21]. The *Conversation Tracker* in turn investigates sequences of posts that refer to each other in order to track conversations about a particular topic throughout the blogosphere. *BlogPulse* is a platform that offers this service for instance[7] . Project *BlogTrace* of *Anjewierden* et al. even goes a step further [22]. By making use of classic mining techniques it aims to investigate the flow of knowledge between blog instances and furthermore tries to look into knowledge-bases of single communities or bloggers. *ThemeSnapshots* by Mei et al. equally allows to track the diffusion of particular topics. It furthermore offers a functionality through which geographic locations can be filtered out, in which particular topics are discussed more intensively than in other regions. *ThemeLifeCycle* is an extension on *ThemeSnapshots* that compares the frequency of post-publications on particular topics in different locations over time [15].All these analyses allow for the investigation of opinion leaders or experts on particular topics in the whole of the blogosphere.

Communities

The structure of linkages allows detecting Communities - also known as blog rings or blog groups - in the blogosphere. Corresponding research projects try to uncover communities of similar interests and opinions. Chin and Chignel for instance examine how communities can be discovered through interconnected blogs as a form of social hypertext. Their proposed model detects communities in blogs by aligning centrality measures from social network analysis with parameters of the "Sense of Community" model (SOC) of Mcmillan and Chavis [24]. A similar approach of Efimoa et al. was noted in the work "Finding the life between the buildings [...]" [24][14]. These can either be blog rings with a similar racist background (also known as hate groups) [25] [11], or blog groups that share a common interest about a particular product innovation [26] [27]. With the help of blog mining methods, statements can be made about a community's structure, opinion leaders within these communities as well as the organization, scale and geographic spread of such blog groups.

[7] http://www.blogpulse.com

Rankings
Blog Mining should also provide some sort of ranking score, through which individual weblogs can be compared with each other. This allows users to find the "best" blog about their particular field of interest. Common ranking algorithms like Google's PageRank [28] or Kleinberg's HITS [29] rank the content of a document (frequency of specific key words) on the one hand, and the number and quality of incoming or outgoing links of that document on the other hand [30]. These metrics were optimized to rank traditional web pages. The application of such ranking algorithms on weblogs would however not be optimal since the number of links between blog posts is not as high as for traditional web pages. Blog specific data like timestamp, author reputation, update rate, different types interlinkages as well as the number of reactions (comments, linkages, etc.) on a certain posting, which was so far not included in traditional ranking algorithms, should definitely be taken into account in a blog-specific ranking algorithm [31][30]. Another approach to rank online content such as RSS feeds, posts or news articles is the one of PostRank. The ranking algorithm of this system is based upon the mode and frequency with which users react on particular online content. Interaction is measured according to five different categories - namely Creating, Critiquing, Chatting, Collecting and Clicking. Creating represents the strongest, Click-ing the weakest form of interaction. The analysis of interaction of a specific post is also influenced by corresponding activity on social platforms like Twitter or del.icio.us.

5.2 Content Analysis

Weblogs are creating a huge amount of unstructured and unfiltered information each day. This section aims to shed some light onto the discussion of how published information in the blogosphere can be filtered in order to conduct meaningful opinion or trend analyses, only to name a few.

Content Filtering
The enormous amount of information in the blogosphere makes it virtually impossible for the individual user to retain only a minimal overview about what is being published about a particular topic of interest. There are several research projects that provide basic content filtering techniques in the blogosphere. The most straightforward methodology for blog analysis is the INSPIRE tool. With this tool, users can harvest, view them by thematic content, isolate key words of interest, run queries, visualize changes in content over time, or isolate bloggers of interest [32]. Other research initiatives include the one of Qamra et al. [33]. This approach tries to find "stories" represented by a set of blog entries that are about a specific issue and that

reflect a discussion in blogspace between members of an online community.

Opinion Detection
The field of opinion detection is somehow related to the area of Community detection discussed before. However, while members of a community can still have different opinions about a specific topic, this analyses focuses on uncovering bloggers with identical opinions - perfectly suitable to analyze trends during political campaigning for instance. Also, within opinion detection, analyses rather focuses on the content of posts, and not so much on the linking structure as in community detection. An interesting real-life example in this regard was for instance undertaken during the 2004 presidential election campaign in the United States, where the political blogosphere of the US was grouped into conservative and democrat supporters [34]. Research in this field of interest was also undertaken by Attardi and Simi [35] as well as Glance et al. [27].

Trend Analysis
Another related form of research within the field of content analysis is the detection of trends or hot topics - also known as "buzz" - currently discussed within the blogosphere. "Vizblog" for instance is a visualization technique to reveal similarities between blog entries in different blogs [36]. Through association and content analysis, blog entries are linked to each other to form clusters of related content. By manipulating the graph and filtering content, their visualization let users to navigate and explore online discussions. It can furthermore promote participation by highlighting 'the buzz' of popular topics (see also [33]) and laying out the structure of conversations (see also [37]). Tirapat et al. have a similar approach in revealing "buzz" in the blogosphere [38]. Their method uses information-retrieval techniques to associate blog entries to topics, collected by crawling an authoritative resource on this subject area. They visualize and systematically analyze several blogs in the same domain of interest (movies) in order to assess to what extent the "buzz" of blogs correlates with public opinion. The collected information is represented in terms of a topic map, which is subsequently visualized in three different types of views, each one designed to communicate a different aspect of the data. A particular service provider that is known for its meaningful trend analysis is BLOGPULSE.

6 Visualization

Part I and Part II (see "Extraction" and "Analysis" in fig. 2) of our long-term research initiative are by now completed. The ultimate step is the implementation of the central user interface, in which the user will be able to explore and assess the data provided by our crawler and analyze it accord-

ing to all the metrics that were previously discussed in section 5. The last project phase "Visualization" presented at the bottom of figure 2 summarizes the user interface as we plan to realize it in the upcoming month in an abstract manner. It is supposed to follow a one-stop-approach, in which the user can obtain all the information needed from one single framework. This visualization is structured along several dimensions, with each one of them visualizing a different abstraction layer of the blogosphere. The top layer of the pyramid-like depiction represents all the information necessary to understand the structure of the blogosphere as a whole. This level thus visualizes most of the information as analyzed in the network analysis as discussed in section 5.1. When zooming into a particular sector of the blogosphere (equivalent to the middle segment of the pyramid-like depiction), the user gets a more detailed view that focuses on partial blogospheres (national blogospheres for instance) or - even more specifically - on single weblogs. The user interface would provide related information like language spoken, author rankings, opinion leaders for particular topics, and linkage visualizations of partial blogospheres in this layer. When zooming even further into a particular weblog of interest, the user would get into the most explicit layer within our user interface (lowest segment of the pyramid-like depiction). Here, users would get blog-specific information that has been analyzed within the content analyses as discussed in section 5.2. Next to the visualization of content- or context related interdependencies within one single weblog, the user interface would furthermore provide detailed blog-information about the topics covered and the most active authors, among many others. There are other rather loosely related projects that are dedicated to the visualization of social networks in general and not the blogosphere in particular: We like to mention MultiVis [39], Viszter [40], Prefuse [41] as well as the work of Shen et al. [42], Trier [43] and Gloor et al. [44] in this regard.

7 Conclusion

Similar efforts are already undertaken by several providers in offering an integrated blog intelligence service. However, there is not one service that offers all those features, and meets all those requirements as the blog intelligence framework presented in this chapter. Even basic metrics as discussed in section five are not incorporated within every platform investigated. Figure one summarizes all those features we believe to be important for such an integrated blog intelligence platform (see sections 5.1 and 5.2). But there are additional requirements on top of these metrics that a decent blog intelligence service platform should meet. We consequently believe that any platform should make the information available that is required for a basic performance-check of their methods employed regarding the collection, the storage and the analysis of their data. For most platforms it was impossi-

ble to retain basic key data like total number of blogs identified, or national blogospheres covered. We also consider it as a major handicap if a blog intelligence service has to rely on data provided by externals, rather than generating their analyses out of their own data archives. According to our research, only BLOGPULSE comes close to an aggregated blog intelligence platform that meets all those requirements we believe to be indispensable for such a service. GOOGLEBLOGS, SPINN3R and AMATOMU do poorly in nearly every evaluation score. POSTRANK, TECHNORATI, ICEROCKET, TWINGLY and BLOGSCOPE exhibit intermingled results that also call for some optimization. Even though we completed the grass-roots tasks of our overall research work with the completion of the project phases I ("Extraction") and II ("Analysis"), there still is some work ahead of us with implementing a visualization tool that is capable of presenting all the data collected by the crawler and processed by the data analyzers. Nevertheless, we did already make considerable progress with the lowest abstraction layer of our visualization part (see figure one). The corresponding and self-developed visualization called POSTCONNECT displays content-related and weblog-specific information. It reached a functional maturity by now, and was subject in an academic paper that is currently under review. BLOGCONNECT, which is currently being implemented, will eventually display network-related dependencies of the blogosphere and will thus represent the top abstraction layer of the visualization part. We believe that the framework outlined in this paper will be unrivaled at the time it is fully completed, since it is the only one that will incorporate all those measures and metrics indispensable for an integrated blog intelligence service. We reckon to publish a first prototype of BLOGINTELLIGENCE by the end of this year.

References

1. S. Herring, L. Scheidt, S. Bonus, and E. Wright, "Bridging the gap: A genre analysis of weblogs," Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS'04), 2004.
2. T. Smith, "Power to the People: Social Media Tracker Wave 3," Retrieved on September, vol. 2, 2008.
3. Technorati.com, "State of the Blogosphere 2009," October 19, 2009 at 6:00 am.
4. J.F. Bross, A.E. Acar, P. Schilf, and C. Meinel, "Spurring Design Thinking through Educational Weblogging," 2009 International Conference on Computational Science and Engineering, Ieee, 2009, pp. 903-908.
5. H. Kircher, "Web 2.0 - Plattform f ́ur Innovation," it - Information Technology, vol. 49, 2007, pp. 63-65.
6. J. Schmidt, Weblogs: eine kommunikationssoziologische Studie, Uvk, 2006.
7. J.F. Bross, M. Quasthoff, S.M. Niven, and C. Meinel, "Implementing a corporate weblog for SAP."
8. J.F. Bross, M. Quasthoff, P. Berger, P. Hennig, and C. Meinel, "Mapping the blogosphere with rss-feeds," The IEEE 24th International Conference on Advanced Information Networking and Application, Perth, Australia: IEEE, 2010.

9. M. Rauner, "So tickt das Wir," Spiegel Online, 2009, pp. 1-4.
10. J. Schr'oder, "Technorati ist tot, die blogcharts leben.," Deutsche Blogcharts (DBC), 2009.
11. M. Chau, J. Xu, J. Cao, P. Lam, and B. Shiu, "A Blog Mining Framework," IT Professional, vol. 11, 2009, pp. 36-41.
12. C. Marlow, "Audience, structure and authority in the weblog community," Time, pp. 1-9.
13. S. Mintert and C. Leisegang, "Liebes Tagebuch ... Sieben frei verf'ugbare Weblog-Systeme," iX-Archiv, vol. 7, 2008, pp. 42-53.
14. L. Efimova, S. Hendrick, and A. Anjewierden, "Finding'the life between buildings': An approach for defining a weblog community," Internet Research, vol. 6, 2005.
15. Q. Mei, C. Liu, H. Su, and C. Zhai, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," Proceedings of the 15th international conference on World Wide Web - WWW '06, 2006, p. 533.
16. A. Aschenbrenner and S. Miksch, "Blog Mining in a Corporate Environment," 2005.
17. S. Herring, I. Kouper, J. Paolillo, L. Scheidt, M. Tyworth, P. Welsch, E. Wright, and N. Yu, "Conversations in the blogosphere: An analysis" from the bottom up," Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS'05), Ieee, 2005, pp. 107b-107b.
18. E. Adar and L.A. Adamic, "Tracking Information Epidemics in Blogspace," Web Intelligence, 2005.
19. D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," Proceedings of the 13th international conference on World Wide Web, ACM New York, NY, USA, 2004, p. 491-501.
20. A. Barabasi and E. Bonabeau, "Scale-free networks.," Scientific American, vol. 288, 2003, p. 50-9.
21. S. Arbesman, "The Memespread Project: An Initial Analysis of the contagious Nature of Information in Online Networks," 2004, pp. 1-9.
22. A. Anjewierden, R. de Hoog, R. Brussee, and L. Efimova, "Detecting knowledge flows in weblogs, 13th," International Conference on Conceptual Structures (ICCS, 2005, pp. 1-12.
23. Q. Mei, C. Liu, H. Su, and C. Zhai, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," Proceedings of the 15th international conference on World Wide Web - WWW '06, 2006, p. 533.
24. A. Chin and M. Chignell, "A social hypertext model for finding community in blogs," Conference on Hypertext and Hypermedia, 2006.
25. M. Chau and J. Xu, "Mining communities and their relationships in blogs: A study of online hate groups," International Journal of Human-Computer Studies, vol. 65, 2007, pp. 57-70.
26. M. Chau and J. Xu, "Studying Customer Groups from Blogs," fbe.hku.hk, pp. 850-856.
27. N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo, "Analyzing online discussion for marketing intelligence," Special interest tracks and posters of the 14th international conference on World Wide Web - WWW '05, 2005, p. 1172.
28. S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," Computer networks and ISDN systems, vol. 30, 1998, p. 107-117.
29. J. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of the ACM (JACM), vol. 46, 1999, p. 604-632.
30. B. Ulicnya, K. Baclawskia, and A. Magnusb, "New metrics for blog mining," au.af.mil, 2007.
31. A. Kritikopoulos, M. Sideri, and I. Varlamis, "BLOGRANK: Ranking on the blogosphere," Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007), Boulder, Colorado, USA: 2007, pp. 2-3.

32. M. Gregory, D. Payne, D. McColgin, N. Cramer, and D. Love, "Visual Analysis of Weblog Content," ICWSM'2007, Boulder, Colorado, USA: PNNL-SA-54063, International Conference on Weblogs and Social Media'07, Boulder, CO, United States (US)., 2007.
33. A. Qamra, B. Tseng, and E. Chang, "Mining blog stories using community-based and temporal clustering," Proceedings of the 15th ACM international conference on Information and knowledge management, ACM, 2006, p. 67.
34. L. Adamic and N. Glance, "The political blogosphere and the 2004 US election: divided they blog," Proceedings of the 3rd international workshop on Link discovery, ACM, 2005, p. 43.
35. G. Attardi and M. Simi, "Blog Mining through Opinionated Words," pp. 2-7.
36. C. Tauro, M.A. Pérez-Quiñones, P. Isenhour, S. Ahuja, and A. Kavanaugh, "VizBlog: Discovering Conversations in the Blogosphere," Technology demonstration at Directions and Implications of Advanced Computing-Conference on Online Deliberation, University of California, Berkeley, 2008.
37. F. Viégas and J. Donath, "Social network visualization: Can we go beyond the graph," Workshop on Social Networks, CSCW, Citeseer, 2004, p. 6-10.
38. T. Tirapat, C. Espiritu, and E. Stroulia, "Taking the Community's Pulse, one Blog at a Time," cs.ualberta.ca, Palo Alto, California, USA: ACM New York, NY, USA, 2006, pp. 169-176.
39. J. Sun, S. Papadimitriou, C. Lin, N. Cao, S. Liu, and W. Qian, "Multivis: Content-based social network exploration through multi-way visual analysis," Proc. SDM, 2009, p. 1063-1074.
40. J. Heer and D. Boyd, "Vizster: Visualizing online social networks," Proceedings of the 2005 IEEE Symposium on Information Visualization, 2005, p. 33-40.
41. J. Heer, S. Card, and J. Landay, "Prefuse: a toolkit for interactive information visualization," Proceedings of the SIGCHI conference on Human factors in computing systems, ACM New York, NY, USA, 2005, p. 421-430.
42. Z. Shen, K. Ma, and T. Eliassi-Rad, "Visual analysis of large heterogeneous social networks by semantic and structural abstraction.," IEEE transactions on visualization and computer graphics, vol. 12, pp. 1427-39.
43. M. Trier, "Towards a Social Network Intelligence Tool for visual Analysis of Virtual Communication Networks," Virtuelle Organisationen und Neue Medien, K. Meissner and M. Engelien, Dresden: TUDpress Verlag, 2006, pp. 331-342.
44. P. Gloor, R. Laubacher, Y. Zhao, and S.B. Dynes, "Temporal visualization and analysis of social networks," NAACSOS Conference, June, Citeseer, 2004, p. 27-29.

Twitmographics: Learning the Emergent Properties of the Twitter Community

Marc Cheong and Vincent Lee

Abstract This paper presents a framework for discovery of the emergent properties of users of the Twitter microblogging platform. The novelty of our methodology is the use of machine-learning methods to deduce user demographic information and online usage patterns and habits not readily apparent from the raw messages posted on Twitter. This is different from existing social network analysis performed on de facto social networks such as Facebook, in the sense that we use publicly available metadata from Twitter messages to explore the inherent characteristics about different segments of the Twitter community, in a simple yet effective manner. Our framework is coupled with the self-organizing map visualization method, and tested on a corpus of messages which deal with issues of socio politi-cal and economic impact, to gain insight into the properties of human interaction via Twitter as a medium for computer-mediated self-expression.

1 Introduction

Twitter [1] is a microblogging platform that allows users to publish status updates in 140 characters or less. Twitter users are using it for reasons as mundane as publishing their thoughts and daily activities, to sharing opinions, ideas and news, and communicating with friends and family [2]. Its

Marc Cheong,
Clayton School of Information Technology,
Monash University, Victoria, 3800 Australia.
e-mail: marc.cheong@infotech.monash.edu.au

Vincent Lee,
Clayton School of Information Technology,
Monash University, Victoria, 3800 Australia.
e-mail: vincent.lee@infotech.monash.edu.au

user base is an amalgam of different age groups, transcending international borders, social status, languages and professions.

More recently, Twitter has been used more creatively in helping 'push the message across', creating a surge in popularity among the service which is already enjoying an annual growth rate of several hundred percent in terms of users [2]. Examples of the expanding realm of Twitter usage range from product marketing and generating awareness such as election campaigns [3] and the Earth Hour environmental campaign [4]; to more extreme cases such as rallying diplomatic help in rescuing a journalist from arrest [5] and covering the aftermath of the 2009 Iran general elections amidst media blackouts and rioting [6].

In other words, the examples above illustrate that Twitter can be considered "a hybrid between communication and social networking" suitable for media mining, based on the findings of several studies [7, 8, 9]. This lends credence to the fact that it can very well be a source of collective intelligence or the 'wisdom of the crowds' [10] which has vast potential to be tapped for gathering opinions and information for effective decision making [11]. Aside from the domain of Twitter message contents (also known as 'tweets') and chatter, a domain that could provide us with an insight of the collective wisdom of microbloggers is the user domain itself. This content is generated by one of two methods:

1. Twitter itself providing basic demographic information of a user; and
2. The user himself authors such data [2, 1].

Such information is exposed by the Twitter Application Programming Interface (API), and simple programs to extract this data already exist (which are relatively simple to implement) and have already been in use in academic research [7, 8, 12, 13].

We have created a framework to discover user demography, habits, and sentiments when contributing to a certain thread of discussion (such as Twitter 'trending topics' or 'hashtagged' messages). Our research is novel as to our knowledge, there is no prior work done on automated discovery of the latent properties of a Twitter communities contributing certain topics.

Our paper is divided as follows: first we cover related state-of-the art in Section 1, then Section 2 details our framework and the algorithms involved. Section 3 is an investigation into the accuracy of the synthesized attributes by comparing the automated procedures to manual human checking, followed by case studies in Section 4 and the paper's conclusion.

2 Prior Work

Several studies have been conducted to study the dynamics of the Twitter community. User intentions and the main style of writing Twitter messages

have been studied by Mischaud [9] and Java et al. [8]. The dynamics of communication of Twitter users in the form of replying habits have been performed by Honeycutt et al. [12]. Studies on the emergent properties of Twitter have been conducted by Krishnamurthy et al. [13], Huberman et al. [14], and also mentioned in [8]; findings from [12, 8, 13] mainly cover the aspect of the social networking pattern exhibited by Twitter users. With regards to HCI, this can be seen as an extension of the human social network dynamic via Twitter.

There have also been several studies done in relation to visualizing online chatter as an extension of real-world communication and interaction. An interesting piece of work by Harris & Kapvar [15], the 'We Feel Fine' project, is an electronic art project which harvests demographic data from newly created blog posts and visualizes it according to the sentiments or emotions of the blog authors: in their own words, creating an "exploration of human emotion on a global scale"[15]. Hazlewood and Makice [16] have developed Twitterspace, which projected the 'tweets' from members of a community on a visual display, to "[blend] the virtual space of Twitter with [their] physical community centers".

Twitter has also been studied as a facilitator in rallying people toward a collective call of action in real life. Characteristics of chatter among Twitter users in times of mass convergence and emergency have been studied by Hughes & Palen [17]. On a related note, research has also been done to study the use of Twitter in the humanities, such as a facilitator for activism [18] and political awareness [19].

The motivation of people to participate in Twitter and similar online social networks has been discussed by Makice [20], where he proposed the use of Twitter in phatic communication in the context of "strengthening community". From a social information perspective, Dearman et al. [21] performed a study on the needs of information sharing and concluded that people have differing needs for information sharing and satisfying their own need for information, and therefore tend to leverage an existing social network to do so. Joinson [22] has studied the role intentions of users on Facebook, a de facto social network not dissimilar to that of Twitter. In a similar vein, Schrammel [23] has performed research on the need for information sharing in an online social network. In the office/work environment, Zhao & Rosson [24] found that Twitter is being used my employees for updating others on their daily goings-on, hence promoting connectedness, expert-seeking, and improving understanding of colleagues.

Cheong & Lee [7] have performed preliminary analysis on the emergent properties of people who choose Twitter as a communication platform. They have studied the reason why users contribute to a certain 'trending topic' of discussion based on their demographic and habitual properties. An interesting point to note is that their habits of communication via computer-mediated social networks can be connected to their core demography [7].

On a related note, memetic activity, similar to what is studied on Twitter by [7], has been conducted for instance on news sites by Leskovec et al. [25], websites and emails [26, 27]; concepts such as memetic tracking and viral information spread is also relevant to studying topic-based properties in Twitter users and messages.

Some key differences with our framework that differ from existing approaches discussed in this section can be summarized as follows. This study automates the harvesting and interpretation of data on Twitter to provide us with a useful set of summary information about users and messages, focusing on one particular topic at a time (intra-topic), rather than across topics (inter-topic); also, our framework automates the task, requiring little manual intervention or post-processing. Existing studies such as [7, 14, 8, 13] are exploratory studies on Twitter users and messages in a more general perspective, and tend to focus more on either the message aspect or the user aspect, but rarely in tandem with one another. Finally, our framework is built upon existing studies from various disciplines, e.g. Twitter in disaster and crisis [17], social information [21], and user intent [22], in order to have a generalized idea of what kinds of useful information that can be gleamed from a microblog perspective.

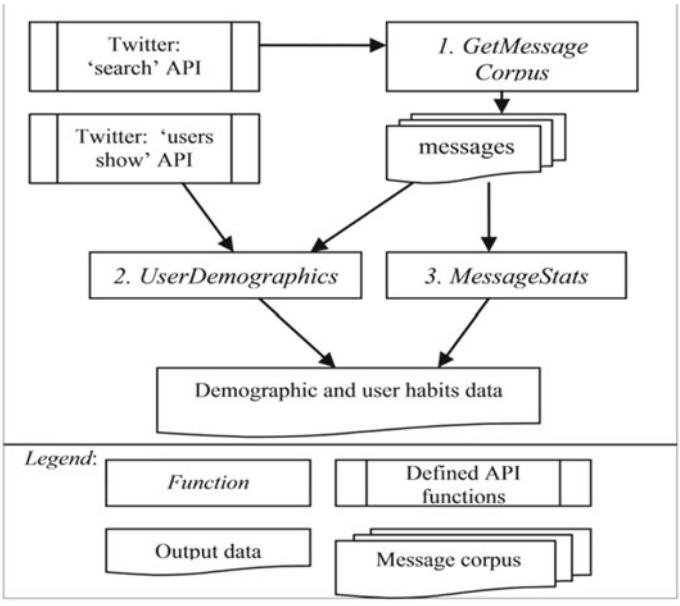


Fig. 1: The Proposed framework.

3 Proposed Framework

The framework for our system is briefly illustrated in Fig. 1. We use Perl to develop our Twitter analysis framework as it is well-suited to processing large chunks of textual-information. This framework can be divided into 3 different processing modules.

1. *GetMessageCorpus* queries the search API for mentions of a topic and performs in-memory hashing and dumping the results to disk for further analysis (while discarding any results that have been analysed during a previous run). The raw data obtained from this module has little value to us; further data has to be derived and generated in the second and third parts of our Perl script, *MessageStats* and *UserDemographics*. Table 1 briefly describes the functionality behind this module.
2. *MessageStats* reveals message statistics, embedded in the metadata of a Twitter message. This will be explained in the next subsection.
3. Finally, *UserDemographics* provide us with the underlying emergent properties from the user base. The user base is based on the authors of the messages harvested in *GetMessageCorpus*.

3.1 Features of MessageStats

The overall MessageStats algorithm returns 7 attributes, and can be illustrated via the algorithm in Table 2.

Device/platform classification classifies Twitter messages as generated by 6 distinct classes of devices and software clients - the official Twitter web interface, mobile devices, social media applications, alternative Twitter software clients, feed aggregators, Twitter 'mashups' for information sharing, and Twitter marketing and bulk messaging tools. Our pool of 'client IDs' contains 66 unique pieces of software; the device and platform are ascertained by the software authors' descriptions in their. This extends the categorization performed in prior work [7, 13] to give a clearer overview of the various techniques users contribute to the 'Twittersphere'. Potential applications of this piece of data include detecting whether censorship has occurred (example in [6]), or to determine the reach of HCI artefacts among Twitter users, for example mobile and ubiquitous computing as opposed to computer-based usage of the microblogging platform.

Message length is a common metric [28] in investigating the sociolinguistic properties of a particular demographic of users. This identifies, for a certain topic, if users use Twitter to broadcast long postings (akin to conventional blogs) as opposed to broadcasting short snippets (summoning help or announcing breaking news on the spot) for example. This reflects the user

Table 1: GetMessageCorpus algorithm.

```
GetMessageCorpus(topic):
        query for topic via SearchAPI
            foreach message in results
            cache results in memory
            dump results to disk
endfor
```

behaviour of information sharing, that can be tied to themes as discussed in [21, 22, 23].

Table 2: MessageStats algorithm.

```
MessageStats(messagecorpus):
    foreach message in corpus
        analyze sources by matching device/platform with software name
        analyze content by length via string length parsing
        analyze content by features via string matching/regular expressions

    UserDemographics(message)
endfor
```

Three methods of content analysis (Reply, Retweet, Hashtag): content analysis is also performed on the raw message data created by Twitter users. Twitter messages frequently contain specially-formatted words that can be picked out to determine the characteristics of the reply:

1. Reply messages: '@user' to indicate a reply to user.
2. Retweets: 'RT message' to indicate 'retweeting' or forwarding on a message
3. Hashtagging: '#keyword' to 'hashtag' a message with keyword

Such kinds of messages could be singled out and categorized to give an overview of how a particular topic of interest rapidly gains popularity [7], how such messages get disseminated via directed replies to users in a particular community [12, 14, 17] and enables us to identify certain properties of a topic with applications in memetics [26] and spike detection [29]. It is also postulated in blog research such as [30, 31] that content analysis similar to what we are performing is able to provide a guess as to the demographics of the contributing user.

Presence of URLs: Presence of URLs indicates an intention by the user to share information (not unlike behaviour exhibited on de facto social net-

works and news aggregators).

Picture attachments: presence of the term 'twitpic' which is a photo-sharing service [32]to indicate the presence of linked images in Twitter message contents; the usage of this is an indication of user-generated content sharing and reporting of eyewitness news. Again, we could refer to the presence of these details as an indicator of a particular user (or community's) need for computer-mediated information sharing [21, 17, 23].

3.2 Features of UserDemographics

UserDemographics returns a set of 8 attributes generated by analyzing the user account of message authors. Its basic operation is illustrated in Table 2.

Table 3: UserDemographics algorithm.

```
UserDemographics(message):
    get username from message
    if username has been analyzed from a previous message
      retrieve results from hash
      return
    query for username via UserAPI
    if username has been banned based on UserAPI return values
      flag as banned
      return

    determine gender via our ranking algorithm
    determine country via Google Geocoder API

    classify web usage habits via string-processing certain API values
    classify Twitter usage habits via mathematical operations from raw API data

    cache results in memory in hash table
    dump results to disk
```

Gender: In prior literature [33], gender is identified as one of the attributes that can be "subtle cues to user identity". Previous research [7] has identified that users on Twitter frequently publish their names as opposed to a 'handle' or nickname as part of their user information. Work in [22, 23] have identified gender as one of the differentiating factors in influencing information sharing and social networking, which we attempt to leverage in our framework. To identify the gender of a Twitter user, we use a simple ranking algorithm to match his/her given (or middle) name the United States Census

statistics on 5494 unique and ethnically-diverse first names.

Location: Twitter provides two pieces of data to identify a users' location. The time zone field and location field can be edited by the user to indicate their location. The usage of GPS-enabled mobile clients allows the user to be tagged with a GPS coordinate, signifying their precise location. Previous works [8, 13] have used the GPS coordinate and the time zone as means to determine the user's country. However, the location field can be populated by names of cities/towns or districts (for example 'Morwell, Australia' or 'Ibaraki Prefecture'), which can be time consuming to map to specific countries but are meaningful in the absence of GPS data. In our approach, we use the Google Maps Geocoder API to map both GPS coordinates as well as query for the country in which a town/city/district resides. In our study, we attempt to differentiate the habits of Twitter users based on geographic location, which plays a role in understanding the reach of online social networks such as Twitter based on the geographic location of the users [34]. The justifications behind using this method over several existing ones include:

- Google has a comprehensive Geocoding API which is programmer-friendly and has an extensive set of location names used by their Maps application. This improves on existing ideas such as usage of Yahoo Geocoding API only on GPS coordinates [8], as Google's API can also handle place names well.
- Existing methods (e.g. from [13]) use the location information based on user time zone, which can be inaccurate; for instance, if the user sets the time zone wrong on his computer.

Web usage habits: the Twitter profile page for a user also lets the user publish the URL of a webpage. Frequently, users publish the URL to their profiles on other social networks (such as Facebook), blogs or content sites in which they share their content (such as YouTube, Flickr, and Wordpress). We filter out and categorize the users' URLs based on these specific patterns; this attribute is useful to determine any connections between a stratum of users and their corresponding social media usage and information-sharing properties. The findings from [21, 17, 23] apply here, as the Twitter profile URL is a publicly available way of pushing forth a user's identity and allows us to gleam an insight into the user's online persona.

Twitter usage habit via profile picture (avatar) presence: In the studies by Nowak & Rauh [35], they note that the presence of an avatar picture via a computer-mediated medium, such pictures help in "identifying, recognizing, and evaluating others in the mediated world of geographically distant communication"; a sentiment which is also shared by Erickson [36] in terms of user visibility. In our interpretation, Twitter users who choose to put profile/avatar pictures are more likely to interact and participate in Twitter

activity as opposed to those who do not.

Friend/follower ratio: this indicates the proportion of users that are followed versus those the user are currently following (in social networking theory, their in-degree versus out-degree). The in-degree and out-degree are frequently used statistics (used in [8, 13] and mentioned in [2]) that provide us with the dynamics of networking/interaction for a particular user (or strata of users).

Account age: characterized by the number of days since the user's last posting from the day the user created an account. This allows for identifying new accounts and the degree of 'veterancy' of users; the potential application for this is detecting user loyalty for Twitter as a medium of expressing oneself [17], to detecting opinion spam and 'sockpuppetry' [11].

Message frequency: derived from the number of messages posted divided by the account age (in days). This metric allows us to see the Twitter participation frequency and any undue influence a particular topic might have in increasing this frequency (for example trending behaviour [29] or emergencies [17]).

Violation of terms of service, in terms of banned accounts: another piece of information vital to understanding the Twitter user base is the presence of banned or deactivated accounts due to violation of terms of use, first applied in [7]. Such accounts, although their proportion may be small, are meaningful as we could detect the users that might have polluted the chatter about a particular topic with spam, misleading messages [11]; and in several cases scamming and phishing behaviour [7].

3.3 Message Harvesting Process

We use the Twitter API provided by Twitter Inc. [1], specifically the search API and the REST (Representational State Transfer[1]) users show API. For the purposes of this research, we requested for white listing from Twitter Inc., allowing us to retrieve up to a maximum of 20,000 pieces of unique user information per hour via the REST users API. However, the search limitation of a maximum 1500 messages (or a dynamic backdated range of

[1] In a nutshell, Representational State Transfer refers to the Twitter API which processes the re-quest for user information and returns it in a representational manner, i.e. as an object we can use. The Search API is used to fetch search results matching our intended keywords; it returns more complete results than the Streaming API which provides only a sample of all available tweets, but in a streaming format.

approximately 20 days, whichever first) is applied to the search API; the workaround is to perform several 'search' operations spaced in an interval of approximately 10 minutes between each run, enabling us to reach the capacity of tens of thousands of messages per hour (to correspond with the maximum user data retrieval limit of 20,000 per hour as stated).

4 Validation Results on Synthesized Attributes

The Twitter data mentioned in the methods above, with exception of gender and location, are drawn directly from the Twitter API; hence the validation of such results appears to be trivial. We list these attributes that do not require validation, and our explanation, as follows:

- **Device and platform classification** is based on simply matching software clients to those already described in the Twitter API (ground truth).
- **Message length** and content analyses (**Reply, Retweet, Hashtag, Presence of URLs**, and **Picture attachments**) can easily be determined via string-processing functions in Perl.
- **Web usage habits** and **profile picture (avatar) presence** can also be determined with string-processing functions.
- **Friend/follower ratio, Account age, Message frequency, and Violation of terms of service checks** are simply calculated via mathematical operations from raw API data.

As for the synthesized attributes for gender and location, we require validation on their accuracy compared to manual mapping and determination of such attributes by hand, as the number of potential locations, GPS coordinates, and possible human names are very large. We perform testing by harvesting user information from a list of 1000 random users, and comparing our automated analysis with the ground truth (i.e. manually checking for correctness).

We tested our simple gender detection algorithm over 10 sets of 100 names each. The names are harvested at random from a set of 1000 Twitter messages. For comparison, we performed a blind test by manually determining or guessing the genders of the names in the test set and compared our algorithm's results with the human-determined results. The reason we break down 1000 names into ten sets is to allow ease of checking and comparison, as the process of manually inspecting and classifying 1000 names can be rather laborious and time consuming.

This algorithm involves the use of the US Census ranking data on first names to probabilistically determine the gender based on a user's real name, which employs a hashing algorithm to pre-load all the first names on the census data for both males and females (with linear complexity as each name has to be pre-loaded only once). Subsequently, each check only involves a hash-

lookup operation, which is a constant-time operation (average O(1)), which should not be a problem to detect gender for a large set of users. This, as far as we know, is the first time such detection takes place in terms of sentiment analysis and demographic projection in the domain of microblogging.

Table 4 contains the result of gender prediction over 10 test sets of 100 names each. Averaging the accuracy rates over each of the ten test sets, we obtain an average accuracy rate of 86.6% (or approximately nine out of ten correct). The algorithm works successfully for common names, for example 'John' and 'Jane'. The comparison obtained is based on the underlying assumption that human (manual) detection always represents the ground truth, i.e. detecting a person's gender based on name will be no problem for human testers.

Table 4: Accuracy of gender prediction via the simple ranking algorithm versus manual determination.

Test set	Number of correctly determined genders (out of 100)
1	81
2	88
3	89
4	82
5	90
6	83
7	85
8	92
9	88
10	88

However, several limitations that affect the accuracy of human validation (and by extension, our algorithmic accuracy) have been identified. These include:

1. Usage of rare names: our algorithm is derived from the US Census ranking data for common first names, and as such is not exhaustive; same applies to humans where not all names will be familiar to a human tester.
2. Androgynous names: names such as 'Tracy', 'Kim' and 'Lauren' are applicable for people of both genders; hence the algorithm (and even human testing) is not able to determine the gender accurately.
3. Presence of names in non-human contexts: if human names are present in non-human contexts, e.g. part of an organization's name, the algorithm does not ignore it as a human tester would. An example such a context is the name 'Beverly Cinema' where it is a cinema's name and not bearing the surname 'Cinema'.

As for the geographic location, we perform the tests for the reverse geocoding (GPS coordinate lookup) and place name finder algorithms based on the Google Geocoder API on 10 similar data sets (totaling 1000 user records). For comparison, we manually identify (with the help of atlas websites) the locations of the places and categorize them based on countries. Countries are determined by the 2-character ISO 3166 country codes.

The complexity behind this method is O(n*k), based on the number of items (n) in the dataset, and the complexity of the invoked proprietary Google Geocoding algorithm (k). A weakness behind this is due to the usage of cloud computing (i.e. using a web server across the Internet to process data) speed bottlenecks such as bandwidth and server rate limiting might affect the processing speed if there is a large dataset involved.

Table 5: Accuracy of country matching by calling the Google Geocoder API versus manual inspection.

Test set	Number of correctly determined country codes (out of 100)
1	96
2	90
3	87
4	90
5	92
6	92
7	88
8	93
9	80
10	89

Table 5 lists the findings of our validation testing. An average of 90.7% detection accuracy (9 cases out of a possible 10) was achieved using the implementation of the Google Geocoder API in our algorithm. This is based on assuming that human detection is the ground truth: i.e. the human tester knows exactly where a particular location is in the world, and which country it exactly belongs to. Strong cases would be for GPS coordinates, and full addresses given, where the country will always be determined successfully.

Several cases have been identified, however, where the location matching mechanism becomes weak. This include random place names (e.g. 'somewhere in the world'), multiple locations (e.g. 'London/Paris/Tokyo'), misspellings of locations (e.g. 'N.Y.Cc'), and the ambiguity of locations (e.g. 'Brighton Beach' could be either in the UK or Australia).

5 Case Studies

We test our exploratory framework on several global-and regional-concern trending topics: the 2009 Iran Election issue [6], US President Obama's reaction toward the Iran issue and foreign policy, and the iPhone OS 3.0 software launch, to see the pattern of Twitter interaction by its user base in regard to differing topics in the current affairs, political, and technology. Our framework-generated data is then used for visualization using a Self-Organizing Map to test the efficacy of using our program as a source of generated qualitative data about a stratum of Twitter users. SOM [37] is a powerful technique based on the artificial neural network concept that projects input from multiple-dimension space into 'maps' of 2-dimensions where similar features are located near each other on the map, which is good for visualization.

We apply our framework to reveal the emergent properties behind the Twitter user base expressing their views on the abovementioned topics. The following table (Table 6) summarizes the keywords, topics, and vital statistics of the accumulated corpus of data.

Table 6: Accuracy of country matching by calling the Google Geocoder API versus manual inspection.

Search term	Total messages (excluding bans)	Banned users	Unique users (excluding bans)
Iran Election	4905	0	1953
iPhone	4246	2	3368
Obama	4640	5	3115

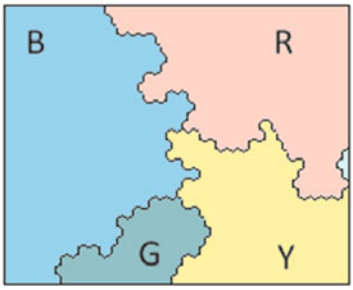


Fig. 2: TSOM clustering of 'Iran Election'.

5.1 Case 1: "Iran Election"

Fig. 2 shows the high-level results of automated clustering and visualization for the term 'Iran Election' (global concern) [6] as fed into a Self-organising Map (SOM) after running the term through our framework. The sentiments of the users discussing this topic can be broken down demographically into 5 clusters.

The blue area (detailed in Fig. 3) represents users from various countries contributing to chatter about Iran's election aftermath. This user base is relatively new, predominantly Iranian web interface users participating on Twitter via the computer as an artefact, with users only adopting Twitter for one month or less, and exhibits an emergent behaviour of frequent reply-based messages, as seen in Fig. 3.

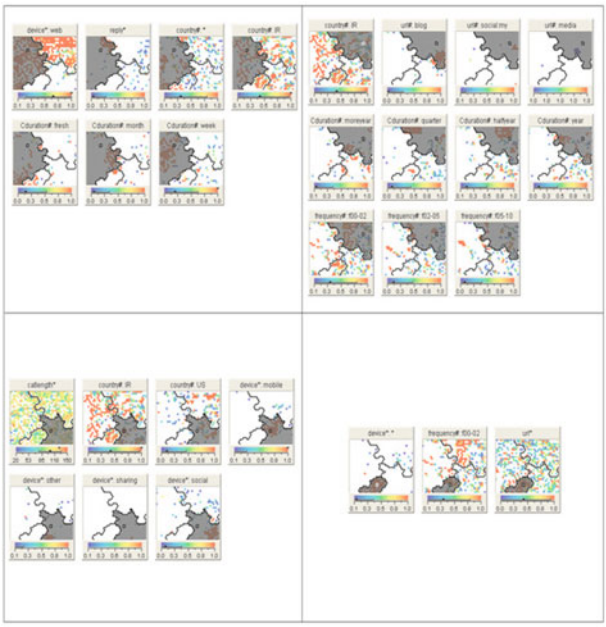


Fig. 3: Emergent attributes for 'Iran Election': **[top-left]** cluster 1/blue, **[top-right]** cluster 2/red, **[bottom-left]** cluster 3/yellow, **[bottom-right]** cluster 4/green.

The red area (detailed in Fig. 3) is made up of almost mainly web users, from Iran and other countries; however this user base is more 'seasoned' or 'veteran' with accounts having been created since at least 3 months. The contribution frequency is predominantly less than 10 messages per day, indicate

sparing usage, but is contrasted by a high usage of other social media sites such as owning a blog or social network page.

The yellow area (detailed in Fig. 3) corresponds to adopters of social media who contribute to Twitter from Iran, the United States and the rest of the world. The inherent features of this cluster can be seen in the usage of mobile devices and social media applications, and the length of messages that hover among the 100-character range. We can deduce from this data that this segment of opinion holders is generating awareness of the Iranian situation via social media, possibly the younger generation.

The green cluster (detailed in Fig. 3) identifies users with little contribution rate (0-2 messages per day), but of varying age of Twitter account, and nationality. The properties emergent from this segment indicate a high posting of URL links in messages, and almost all of them using new categories of Twitter clients that are low in usage. The concentrated use of little-known Twitter clients suggest the depth of the Iranian elections topic in the sense that a very large spread of Twitter users participated in this topic of conversation.

5.2 Case 2: "iPhone"

We now demonstrate the use of our framework to map out demographics for people using Twitter as a medium to express their thoughts about a consumer product (scope being global, the area of discussion pertaining towards marketing and economics). Our framework is tested on the Twitter trending topic 'iPhone' coinciding with the release of a new phone model by Apple Inc.

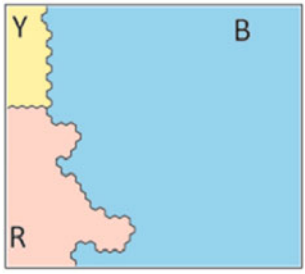


Fig. 4: SOM clustering of 'iPhone'.

Here, we identify three distinct groups with several distinct emergent properties (Fig. 4). The blue cluster (detailed in Fig. 5) identifies the source of the majority of chatter on Twitter regarding the iPhone. The demographics identified from this cluster are male, Twitter 'veterans' who have been adopting

microblogging for a quarter year or more but have an average daily contribution of less than 5. This user base comes from mainly Western countries in which the iPhone has been introduced to market. This user base contributes to Twitter from a variety of devices (mobile and social network-enabled applications inclusive), and also have their own blog/website or social media site.

The second largest cluster (detailed in Fig. 5) is coloured red, whose emergent properties are accounts on Twitter that are significantly new (< 1 week), sourcing data from feeds such as RSS, have a significantly high ratio of followers to 'followees' (higher in-degree than out-degree), and some contribute more than 50 posts per day on Twitter. Most of such messages have URLs in theme suggesting posting of links and content sharing by the users; and the majority of them have no country specified and no gender ascertained which suggest postings by news organizations or news aggregator sites. A small subset of this cluster which is of interest are messages belonging to Japanese (country code JP), which is notable as it reflects rather accurately the market sentiment of the iPhone's new model launch in Japan [38].

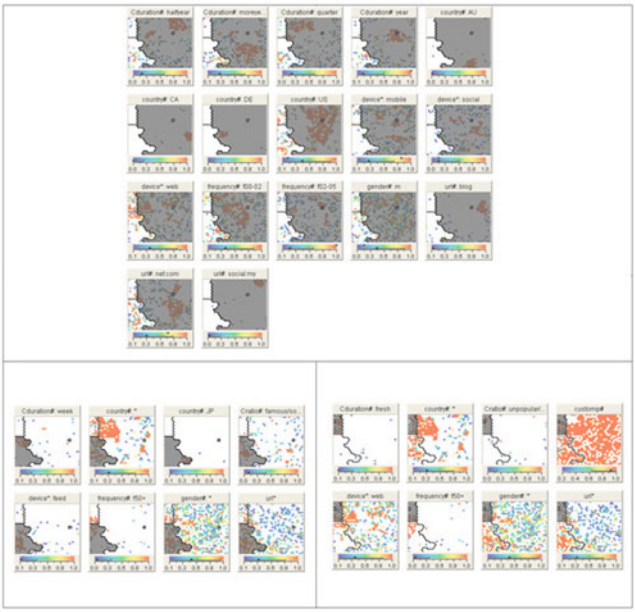


Fig. 5: Emergent attributes for 'iPhone': **[top]** cluster 1/blue, **[bottom-left]** cluster 2/red, **[bottom-right]** cluster 3/yellow.

The final cluster, which is the smallest, is coloured yellow in the SOM above (detailed in Fig. 5). The notable attributes and features for users of this cluster are that they are fresh accounts (1 day old at the most), with un-

popular social connections (following more people than they are friends with), with half of this cluster's Twitter accounts lacking in profile customization. They are posted predominantly from the web interface, frequently have more than 50 messages per day, and mention URLs in the links. As for demography, the gender and location information frequently could not be ascertained at all. We suspect that the Twitter chatter patterns for this group of users reflect those of opinion-spam and 'sockpuppetry' (as hypothesized by Pang & Lee [11]) which pollute the conversation stream with unnecessary noise. In other words, this can be defined as the usage of Twitter not for information sharing and interpersonal communication, but for spamming and other disruptive purposes.

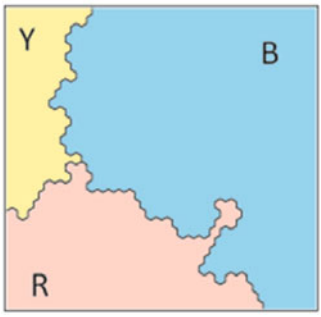


Fig. 6: SOM clustering of 'Obama'.

5.3 Case 3: "Obama"

The final case study conducted is a regional (mostly American) concern which delves into the realm of sociopolitics, with a high-level SOM, visualized in Fig. 6. The keyword 'Obama' (for the US president) is tracked on Twitter to study the impact of his recent foreign policy statements on Twitter user sentiment.

The biggest cluster of the user base detected in this study is the blue cluster (de-tailed in Fig. 7) representing the demographics of Americans naturally concerned about the implications of Obama's foreign policy change. The user base contributes from a large array of devices (social-networking applications, mobile phones), but mainly via the web; a substantial proportion of the users having a website or blog. It is interesting to note that this cluster consists of users genuinely discussing about this topic, as their accounts are mainly more than 3 months old, their messages are almost always long, and their messaging style is focused towards replies, indicating conversation.

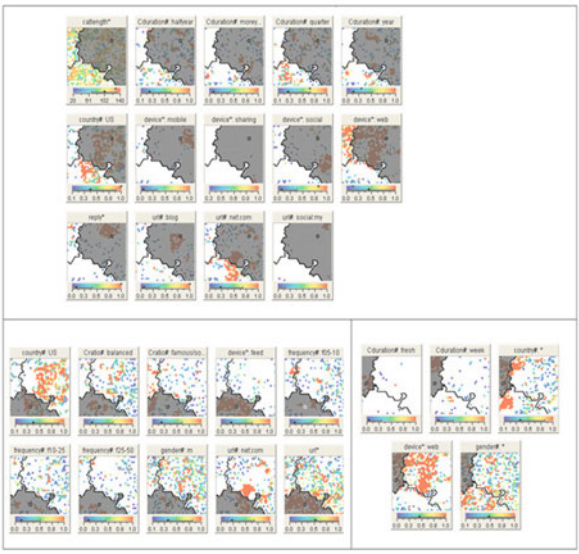


Fig. 7: Emergent attributes for 'Obama': **[top]** cluster 1/blue, **[bottom-left]** cluster 2/red, **[bottom-right]** cluster 3/yellow.

The second largest identified cluster is coloured red (detailed in Fig. 7). We speculate that this cluster belongs to news sources and opinion holders as the demographics reveal that users in this cluster have many followers (refuting the notion of opinion-spamming), predominantly US males, sourcing data from mostly data feeds such as RSS, and frequently publish URL links in their messages.

Finally, we analyse the final cluster coloured yellow (detailed in Fig. 7); this is composed of mainly new accounts, from indiscernible countries and genders, which mostly contribute postings from the web, lending us to suspect the process of opinion-spam as discussed in Section 4.2.

We conclude this case study by reiterating the richness in demographic data harvested from users for socio-political topic areas in sentiment and opinion polling and its potential benefit in demographic segment targeting.

6 Conclusion

In this study, we have come up with a framework in which to perform demographic analysis of users contributing to particular topics on the Twitter microblogging service. We harness the latent demographic data and communication habits of the user base in terms of self-expression and information

sharing via an online microblogging service, to explore the motivations and emergent properties of the users themselves.

We have reviewed how differing habits of users with regards to information sharing and dissemination on Twitter can reveal how users use computers and mobile devices as artefacts in online interaction. A potential application of this is to supplement the traditional techniques of opinion mining, market segmentation and decision making by virtue of its application particularly in the sociopolitical and economic sectors.

References

1. Twitter Inc.: Twitter. Available from http://twitter.com/. (2009) Accessed 16 December 2009.
2. O'Reilly, T., Milstein, S.: The Twitter Book. O'Reilly Media, Inc., Sebastopol, CA (2009)
3. Harris, M.: Barack to the future. Engineering & Technology 3(20) (2008) 25
4. WWF: World Wide Fund for Nature: Earth Hour. Available from http://-www.earthhour.org/. Accessed 16 December 2009. (2009)
5. Simon, M.: Student 'twitters' his way out of Egyptian jail. CNN. Available from http://-www.cnn.com/2008/TECH/04/25/twitter.buck/. (April 25 2008) Accessed 16 December 2009.
6. Fleishman, J.: Mideast hanging on every text and tweet from Iran. Los Angeles Times. Available from http://articles.latimes.com/2009/jun/17/world/fg-iran-image17. (June 17 2009) Accessed 16 December 2009.
7. Cheong, M., Lee, V.: Integrating web-based intelligence retrieval and decision-making from the Twitter Trends knowledge base. In: Proc. CIKM 2009 Co-Located Workshops: SWSM 2009. (2009) 1-8
8. Java, A., Song, X., Finin, T., Tsen, B.: Why we Twitter: An analysis of a microblogging community. In: Proc. 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, Springer-Verlag (2009) 118-138
9. Mischaud, E.: Twitter: Expressions of the whole self. Master's thesis, London School of Economics and Political Science (2007)
10. Surowiecki, J.: The Wisdom of Crowds. Abacus, London (2005)
11. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Volume 2 of Foundation and Trends in Information Retrieval. now Publishers Inc., Hanover, Boston MA (2008)
12. Honeycutt, C., Herring, S.: Beyond microblogging: Conversation and collaboration via Twitter. In: Proc. 42nd Hawaii International Conference on System Sciences. (2009) 1-10
13. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about Twitter. In: Proc. WOSN'08. (2008) 19-24
14. Huberman, B., Romero, D., Wu, F.: Social networks that matter: Twitter under the microscope. Available from http://ssrn.com/abstract=1313405. (2008) Accessed 16 December 2009.
15. Harris, J., Kamvar, S.: We Feel Fine [online media exhibit]. Available from http://-www.wefeelfine.org. (2009) Accessed 16 December 2009.
16. Hazlewood, W., Makice, K., Ryan, W.: Twitterspace: A co-developed display using Twitter to enhance community awareness. In: Proc. Participatory Design Conference. (2008) 230-234
17. Hughes, A., Palen, L.: Twitter adoption and use in mass convergence and emergency events. In: Proc. 6th International ISCRAM Conference. (2009) 248-260

18. Jungherr, A.: The DigiActive guide to Twitter for activism. Available from http://-www.digiactive.org/wp-content/uploads/digiactive_twitter_guide_v1-0.pdf. (2009) Accessed 16 December 2009.
19. Goolsby, R.: Lifting elephants: Twitter and blogging in global perspective. In Liu, H., ed.: Social Computing and Behavioral Modeling. Springer-Verlag (2009) 1-7
20. Makice, K.: Phatics and the design of community. In: Proc. CHI 2009. (2009) 3133-3136
21. Dearman, D., Kellar, M., Truong, K.: An examination of daily information needs and sharing opportunities. In: Proc. CSCW'08. (2008) 679-688
22. Joinson, A.: Looking at, looking up or keeping up with people?: Motives and use of Facebook. In: Proc. CHI 2008. (2008) 1027-1036
23. Schrammel, J., Koffel, C., Tscheligi, M.: How much do you tell? Information disclosure behavior in different types of online communities. In: Proc. 4th International Conference on Communities and Technologies. (2008) 275-284
24. Zhao, D., Rosson, M.: How and why people Twitter: the role that micro-blogging plays in informal communication at work. In: Proc. GROUP'04. (2009) 243-252
25. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proc. KDD 2009. (2009) 497-506
26. Arbesman, S.: The Memespread Project: An initial analysis of the contagious nature of information in social networks. Available from http://www.arbesman.net/-memespread.pdf. (2004) Accessed 16 December 2009.
27. Wasik, B.: And Then There's This: How Stories Live and Die in Viral Culture. Penguin Group (USA), New York, NY (2009)
28. Ling, R.: The sociolinguistics of SMS: An analysis of SMS use by a random sample of Norwegians. In: Computer Supported Cooperative Work. Volume 31 of Mobile Communications. Springer, London (2005) 335-349
29. Gruhl, D., Liben-Nowell, D., Guha, R., Tomkins, A.: Information diffusion through blogspace. In: Proc. WWW 2004. (2004) 491-501
30. Argamon, S., Koppel, M., Pennebaker, J., Schler, J.: Mining the blogosphere: Age, gender and the varieties of self-expression. First Monday 12(9) (2007)
31. Herring, S., Scheidt, L., Bonus, S., Wright, E.: Bridging the gap: A genre analysis of weblogs. In: Proc. 37th Hawaii International Conference on System Sciences. (2004) 1-11
32. Twitpic Inc.: Twitpic - Share photos on Twitter. Available from http://twitpic.com/. (2009) Accessed 16 December 2009.
33. Jones, R., Kumar, R., Pang, B., Tomkins, A.: "I Know What You Did Last Summer" - query logs and user privacy. In: Proc. CIKM 2007. (2007) 909-914
34. Marsden, G.: Using HCI to leverage communication technology. Interactions of the ACM 10(2) (2003) 48-55
35. Nowak, K., Rauh, C.: The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. Journal of Computer-Mediated Communication 11(1) (2005) 48-55
36. Erickson, I.: The translucence of Twitter. In: Proc. Ethnographic Praxis in Industry Conference 2008. (2008) 58-72
37. Kohonen, T.: Self-Organization and Associative Memory. Springer, Berlin (1984)
38. Martin, R.: CNET Asia Blogs: Tokyo Shift - WWDC and the iPhone 3GS. Available at http://asia.cnet.com/blogs/tokyo-shift/post.htm?id=63011359. (June 20 2009) Accessed 16 December 2009.

Dissecting Twitter: A Review on Current Microblogging Research and Lessons from Related Fields

Marc Cheong and Vincent Lee

Abstract Twitter as a microblogging service is fast gaining momentum in the past few years. Publications on the state of the art of various aspects Twitter are summa-rized in this review; which is structured to reflect the different categories of research that can be conducted on Twitter. The bulk of research has been identi-fied to come from the message domain on Twitter, and so far progress has been made to bridge the two domains of Twitter (the message and the user), albeit in a limited form. This review also draws from findings in related fields that could be applied in the field of microblogging, such as research on the 'blogosphere and blog trends; viral information on the web and memetics; and human factors in in-formation sharing and online presence, to name a few. This review provides re-searchers with an insight on to the various problem domains in microblogging re-search, and highlights the links between microblog research and other domains of research. Also, we show that current research rarely bridges the gap between the user and the message domains, and suggest potential improvements.

1 Introduction

Twitter is becoming popular as a data source in, and also a subject by itself, of current research, particularly in the information technology and humanities

Marc Cheong,
Clayton School of Information Technology,
Monash University, Victoria, 3800 Australia.
e-mail: marc.cheong@infotech.monash.edu.au

Vincent Lee,
Clayton School of Information Technology,
Monash University, Victoria, 3800 Australia.
e-mail: vincent.lee@infotech.monash.edu.au

disciplines. Twitter, a Web 2.0 microblogging service, allows users to post 140-character status updates to inform their 'followers' about things ranging from as simple as their current well-being, to sharing the latest social media (be it websites, YouTube videos, and the like). In Twitter's original ethos [1], the point of a 140-character microblog post is for users to answer the question "what are you doing?"[1]

From the first academic paper solely focusing on Twitter which became available circa 2008-2009, it has mushroomed to an anthology of research, some just newly becoming available, even as this paper is prepared. Listings of Twitter and microblogging research document as many as approximately 40 papers within the past two year period (at press time), as documented by danah boyd[2] [2].

Realizing the importance of such research on Twitter, we acknowledge the relevance such academic work focusing on microblogging in our understanding of not only social media and the Web (of which Twitter is an integral part of); but also in such related fields as anthropology, computer-human interaction, data mining, knowledge discovery, visualization - to name a few.

As a result we review several pieces of work, which range from academic papers highlighting Twitter as a subject matter in itself, to books and working papers documenting the human aspect that is visible on Twitter-based interaction, to known implementations or 'mashups' (common parlance in Web 2.0) that extend the existing data on Twitter with other creative uses.

A methodology we use behind classifying papers in this review is identifying the domains of interest current research deals with. We have identified two broad domains in microblogging and Twitter-related research:

1. The user domain: this refers to all the properties exhibited by a user in a microblog setting, which is accessible via the Twitter API. This includes statistics such as 'tweet' count, account age, user customization, and so forth; which allows us to study the human factor behind microblogs.
2. The message domain; this refers to the properties exhibited by a single message composed by a Twitter user. The raw data extractable from Twitter API with this regard includes the message content; software client used, timestamp, geolocation properties, and embedded content.

In the reviews of current work below, we shall frame our analysis and critique based on one or both of the abovementioned domains.

[1] This has changed in late 2009 to "What's happening?" in tandem with the evolution of Twitter's user base who have used Twitter for reasons beyond the original intention of updating others on 'what they are doing now'.

[2] The lowercase form is the author's name used in citations.

2 Exploratory Studies on Twitter

2.1 Twitter's Properties and Emergent Features

The earliest known research conducted specifically in the realm of Twitter is by Krishnamurthy et al. [3] and Java et al. [4]. Java et al.'s pioneering paper in KDD2007 has been expanded into a chapter in [4], which is one of the authoritative citations in Twitter research.

The investigations in Java et al. [4] mostly focused on:

- Approximating the Twitter user growth rate to be ∼1 million User IDs per month, and the message growth rate to be ∼40 million Message UIDs per month (at time of writing, April-May 2007).
- Twitter has the same power law of 'degree distributions' (in-degree versus out-degree in the 'follow' relationship of the user domain) as the Web and conventional blogs.
- Geographic analysis of countries based on the usage of GPS coordinates by users who have clients that use GPS coordinates as 'location' information (user domain).
- Community detection and user categorization is highlighted from a social net-working aspect. Opinion leaders, general users, and information seekers in each 'community' or subset of the user network is determined based on their in-degree versus out-degree and their message frequency.
- Frequent trends in a particular community are extracted by studying the emergent topics in the message domain.
- User intention is studied using an analysis on the message content to determine patterns of chatter, communication using the @user notation, information sharing using URLs, and news reporting by culling RSS feeds.

The research in [4] involves harvesting messages from the public timeline, and then obtaining information in the user domain via the Twitter API involves a crawl of users to obtain the information in the user domain needed to accomplish the above results. For starters, they crawled the public timeline (message domain) by obtaining latest public tweets at intervals, then further obtaining information on the authors of such public messages (user domain), for a total corpus of "1,348,543 posts from 76,177 distinct users" [4]. Social network analysis is then performed on the set of users to detect communities, using graph theory, to compare their findings with standard blogs. Also, they pioneered the use of geocoding to accurately pinpoint the location of users based on their GPS coordinates broad-casted by certain Twitter-enabled devices such as mobile phones. Finally, [4] introduced several categories of user intentions on Twitter, based on their observation on the Twitter communities found in their dataset.

Krishnamurthy et al. [3] on the other hand focused on "distinct classes of ... users and their behaviors, geographic growth patterns, [and current as of August 2008] size of the network". Their message harvesting is based on both analyzing the public timeline (cf. the message domain), and also crawling the users by using the links among out-degree-connected users (the 'follow' relationship in the user domain). Work by Krishnamurthy et al. [3] is similar to Java et al. [4], but using a different methodology and is mainly done in a 'big picture context' compared to [4]. Krishnamurthy et al. [3] conducted the following research in the user domain:

- Characterizing users based on their in-degree versus out-degree ratio.
- Characterizing users based on the client used to publish tweets (limited as of August 2008) and the timestamps of tweets.
- Examining the geographic properties of users via their UTC time zone and their domain name.
- Estimating the size of the entire Twitter user base (at time of writing, August 2008) as approximately 1.4 million.

The investigation in [3] is performed using a similar user/message harvesting and crawling method as [4]. It is done by harvesting the public timeline first; but this time the obtained user data is expanded by subsequently performing a 'crawl' based on the user's friend/follower list. This effectively harvests a range of random user information from the Twitter social network using a constrained crawling method. After acquiring such information, Krishnamurthy and his team worked primarily in the user domain to investigate features of Twitter users.

A technical report by Huberman et al. [5] as of December 2008 has performed an independent analysis into the social networking aspect (user domain) of the Twitter 'follower' network. Their methodology for data acquisition is similar to those of [4, 3], but they focus solely on the social networking aspect of Twitter, by studying the connections in the friend/follow network of Twitter users. They also look up the number of messages posted by a user, and determine correlations between this and the users' social connections. Primarily, this work is framed within the context of the user domain, with limited analysis of the messages, save for the checking of addressivity in the message content. Findings from [5] include:

- Almost 25% of posts are in the form of @user, indicating addressivity of messages to friends/followers.
- The number of tweets composed by a user increase to an asymptotic limit as the number of followers (in-degree) increases; the same holds true for the increasing number of friends (out-degree), but reaching no asymptote.
- Their findings conclude that Twitter users have "two different [social] networks" on Twitter, made up of a "dense one" of followers/friends, and a "sparser" one for actual friends, indicating the need for future research to identify the 'actual' social network of a user on Twitter.

The main area of focus amongst these papers is the user domain, in which the characteristics of users - such as the social networks of a user or communities a user belongs to - are given ample attention and scrutiny. The message domain, however, is only studied to a limited extent.

The following section expands on some of the findings here, in particular the addressivity in Twitter messages dealing with replies (so-called @user messages), and forwarded messages in Twitter-speak ('retweeted' or RT messages).

2.2 Message Addressivity and Forwarding on Twitter

Work in this theme can be attributed to Honeycutt & Herring [6] and boyd, Golder & Lotan [7], who focus on the replying and 'retweeting' (forwarding) aspect of Twitter messages respectively. Such practices of messaging are found in current methods of communication such as email; [7, 6] provides an explanation into such practices in the domain of Twitter and microblogs.

Based on the user intention of 'communication' by [4], Honeycutt & Herring [6] performed research into the usage of such @user reply messages, and into the existence of coherent conversation patterns among Twitter users. Their study primarily focuses on the message domain, using the message properties of 'from user' and 'to user' to identify threads of conversation using the @user messaging pattern. Again, the data harvesting method used is similar to that of the previous section (Section 2.1), by observing the public timeline for a set of random messages, then querying the Twitter API to obtain user information about their authors. Content analysis to determine the particular usage of the '@' symbol is performed, and then the authors employ methods such as dynamic topic analysis to study the coherence of 'chatter' among users via the @user notation.

Honeycutt & Herring [6] concludes by noting that approximately 91% of messages with a '@' sign are intended to signal correspondence between users, and suggests that although Twitter was originally meant to be used to publish status updates, it can indeed be adopted as a platform for purposes of conversation and collaboration. It is important to note that [6] is one of the first papers (in early 2009) to research on Twitter as a platform for conversation, and that their findings are vindicated by the high usage of Twitter for interpersonal communication as of time of writing this paper (end of 2009), as can be seen in Boyd et al. [7].

Boyd et al. [7] expanded upon [6] by studying the 'conversational aspect' of retweeting (in the message domain, detailing the forwarding of messages). The stated goals of the paper are to "describe and map out [retweeting] conventions... [and] examine retweeting practices" and also draw a similarity between "link-based [conventional] blogging". Their data collection method is similar to the ones above (involving the public timeline), but with the

added secondary investigation of directly searching retweets via the Search API. Content analysis on retweeted messages is performed by string-matching for indicators of user intention such as information sharing via URLs in the Retweeted text. This study is conducted primarily in the message domain, where the user domain is just used to provide context to the flow of information between people in the chain of retweets. A summary of findings follow.

- Retweeting has no common convention (as of time of writing, as a new Ret-weeting feature in the Twitter API was only introduced to Twitter near the end of 2009).
- Retweeted messages have elements of information sharing and social tagging, as in the presence of URLs and hashtags. Cascading retweets (like cascading forwards in email) are also common.
- The main motivations of retweeting include spreading tweets, to start a conver-sation, and to draw attention to the originating user.
- Collective action (e.g. to promote awareness, or to use 'crowdsourcing' in find-ing answers to problems) are also a motivation for users to retweet messages.

This paper complements the findings of [6] by incorporating elements of message retweeting in conjunction with studies on message addressivity and conversation. In a way, the study of retweeting is similar to that in [6], but by studying retweet-tagged (e.g. 'RT') rather than the '@' notation. It is pertinent to note that discussions on these two facets mainly involves studying the message domain on Twitter, as the message content reveals more information than studies conducted on users themselves.

3 Information Spread and Self-organization on Twitter and Related Disciplines

In addition to small-scale conversation and retweeting, literature regarding wide-scale information spread and self-organizing behavior among users have also been covered.

3.1 Twitter in Crisis and Convergence

Sutton, Palen & Shklovski [8] first came up with studies of backchannel communication - defined as "public peer-to-peer communication" - in analyzing the use of social media during the 2007 Southern California Wildfires. Their findings pin-pointed the growth and efficacy of social media channels in disaster and crisis response by the general public.

This idea is expanded upon by Hughes & Palen [9] in studying the adoption and use of Twitter in "mass convergence and emergency events". By studying the Twitter activity in the Democratic and Republican National Conventions (political convergence events), and Hurricanes Gustav and Ike in 2008, they observe the usage patterns of Twitter during such events and the type of information shared in the form of tweets.

Hughes & Palen [9] first harvested messages detailing the aforementioned mass convergence and emergency events via the Search API using search terms related to the above events; the selection of search terms was by trial-and-error. After obtaining a corpus of such messages, they studied the message domain to tally the frequency of message posting prior to, during, and after the event. Next, they determine the average frequency of messages per user, and their total contribution to the chatter regarding the aforementioned topics. The conclusions drawn from this first phase of analysis is that the overall per-event tweet count "corresponds with the size and impact of each event", and that a unique user contributes less than three tweets when discussing about an event in general.

Next, the authors [9] analyzed the proportion of reply tweets (in the format @user) and URL sharing in tweets which is an indicator of information sharing and interpersonal communication. Their results show that the proportion of reply tweets is generally less than average (as users may be prone to using Twitter to broadcast information and not for direct communication in these cases); and that URL tweets are much higher than average, indicating that such events have a "higher information demand".

Moving onto the aspect of authors (user domain) of the tweets, the authors in [9] determined that new users joining Twitter in the wake of mass convergence and emergency events tend to adopt Twitter use in the long term, in contrast with the general population on Twitter.

A follow up research by Starbird et al. [10] on the Canadian Red River Valley floods of 2009 studied the "social life" of Twitter messages and the self-organizing behavior exhibited by users discussing the floods. Their data-acquisition methods are similar to that in [9]; however their data processing is significantly different with the added bonus of using a visualization tool designed for easy analysis of data sets (the e-data viewer tool). However, as opposed to [9] who primarily studied the message domain, [10] delved into the user domain by using their visualization tool to facilitate exploring user profiles and code qualitative data such as author locations and source/device used to publish the tweets, as well as potential affiliations to organizations such as media providers.

In the message domain, the authors [10] study behaviors observed in the messages generated by the users from the previous observation. They found many indicators in the messages, among them: "commentary and the sharing of higher-level information" [10]; reply and URL sharing behavior mentioned earlier in [9]; sharing of experiences among flood survivors; and combination

of tweets with authoritative news sources [10] are exhibited in their research sample.

3.2 Pattern Detection and User/Message Clustering on Twitter

In Cheong & Lee [11], the idea of using user and message properties (pertaining to tweets of a certain subject) to segment the pool of messages and their contributors is introduced. This allows for easy visualization when coupled with an existing clustering algorithm (e.g. Kohonen self-organizing maps), allowing us to see the different 'communities' discussing on a particular topic by only depending on the user demographic properties and messaging styles exhibited. Cheong & Lee [11] performed manual analysis and interpretation of the user base: gender is manually deduced from the user's real name, while device used to compose tweets is manually deduced from the name of the software used (cross-referencing with the Twitter API documentation). Primary usage pattern and geographic location are also manually deduced by observing the user profile page by hand.

Besides the clustering methodology employed, they also perform a simple timeline analysis on trending Twitter topics [11] to classify the spikes into 3 based on the observation time window the Twitter API provides (based in turn on enforced API-call limitations).

Related to the timeline analysis referred to in [11], Cheong & Lee [12] also used Twitter spatiotemporal data to track the spread and measure efficacy of a Web 2.0-based activism campaign, i.e. the Earth Hour 2009 campaign in Australia. By measuring state-by-state mentions of Earth Hour activity, Cheong & Lee was able to obtain an indicator representative of each state's participation in Earth Hour by time zone.

Finally, a related technical report by Cheong [13] describes an experiment where all Twitter trends are harvested over a period of ten days, and then analyzed to determine the popular topics in everyday Twitter chatter. Focusing on the textual content, without making any assumption on the underlying user base - in other words, a 'high-level, big-picture' survey [13] - Cheong was able to infer several properties related to the user base contributing toward discussion on a particular trend. By looking at the textual content of the trend, one can trivially deduce, for example Twitter chatter on a Korean female artiste will naturally reveal the user base contributing to this topic as Korean Twitter users who are female.

3.3 Related Literature: On Viral Information Spread and Memetics

Literature on viral information spread (though not specific to Twitter) has also been given attention, as patterns of viral information spread is observed from literature mentioned above.

The spread of viral information (memes) on the Internet has been mentioned as far back as 2000 [14], where memes that spread across the Web and via email have been known to exist before the days of social media. A 2004 study on memetic spread performed by Arbesman [15] on blogs show that the spread of viral information can be observed in terms of relative 'spikes' or increases in traffic to the memetic blog in question. By creating a sample memetic blog, and linking it to just one heavily-trafficked blog, the author [15] found that the memetic blog in question spread like wildfire, causing a heavy spike in traffic, thereby exhibiting memetic behavior within one day of publication. The experiment in [15] illustrates the power of the Internet in its ability to spread information similar to an "epidemic".

Wasik [16] has documented experiments on memetic spread of information (via email, blogs and the Web) conducted from 2003-2004. The memetic spread of ideas in collective action (flash mobs), entertainment, politics, and corporate marketing are revealed in a series of experiments which are well-documented. Examples of such are spreading a 'viral' (in the memetic sense) blog rallying people to a fake cause, and observing the real-world consequences of such a rally. A parallel can be drawn from the 'tipping points' theory posited by Gladwell [17] and the 'wisdom of crowds' theory by Surowiecki [18] which illustrate collective coordination in helping spread an idea until it reaches 'critical mass'.

Such memetic activity is no stranger to Twitter, where Twitter memes such as "#followfriday" (where users list the names of other users who are interesting to 'follow' on Twitter each Friday) and "#musicmonday" (users recommending the music that they are currently like, on every Monday). In fact, a list of such Twitter memes have been listed in crowd-sourced websites such as WhatTheTrend [19] which allows users to assign meanings and interpret such memetic trends for the reference of others. News sites such as The Independent [20] also dedicate a section to such Twitter memes and trends, explaining their prevalence in common everyday Internet usage.

3.4 Related Literature: Approaches to Trend Analysis from Existing Blogs and Social Media

Event detection and trend analysis are common research areas in blogs and other social media that are of interest to us in the study of Twitter and microblogging.

We now review literature related to using online social media and blogs to 'mirror' real world happenings. Gruhl et al. [21] in 2004 studied blogs in an attempt to predict real-world rankings of books on Amazon.com based on the volume of chatter generated in the blogosphere. By acquiring blog posts fed through IBM's WebFountain project, a large archive of blog postings, and then inspecting Amazon's sales rank data, they perform time series analysis to detect any correlations between the two data sources in cases involving mentions of books (c.f. Amazon) by blogs (c.f. WebFountain). What they found is that there is a correlation between the 'buzz' generated due to high mentions of chatter on a particular book with the real-world sales performance on Amazon.com in terms of sales rankings. Their work also introduced a novel concept of spike prediction based on the knowledge of existing blog chatter of a particular book, and discuss the notion of 'spikes' in terms of sales rank.

Choudhury et al. [22] performed a similar study in 2008 to correlate blog communication dynamics on particular stock counters with their real-world performance in the stock market. They use SVM regression techniques to predict the actual movement for a stock, and compare it against the real world activity in the stock exchange. Their framework managed to map the behavior of blog commentary with the real-world performance of a particular stock, with a low error rate. The papers [21, 22] suggest that online chatter on blogs can efficiently be a 'mirror' of real-world happenings.

Fukuhara, Murayama, and Nishida [23] also found a link between blog articles with 'real-world temporal data', where mentions of topics in the Japanese blogosphere are found to have a connection to real-world social events, weather, and topics reported in the Japanese mass media. They employ scripts to harvest blog data and detect spikes using the relative number of mentions of a particular subject over time. Next, they also introduce several patterns of social concerns by observing these time-series graphs. A novel way of mapping real-world temporal data is illustrated by matching the 'social concern' of temperature changes by comparing blog mentions (and spikes) to real-world weather data. Gruhl et al. [24] also concluded that 'spike topics' from real-world events can affect 'spiking' behavior in blog postings.

As mentioned earlier (e.g. in [12]), ideas such spike/trend analysis from the existing blogosphere and social media can be easily ported to the domain of study in Twitter and microblogs in general. The ideas described in this section can be used albeit with a little modification in any related studies of spikes and timeline trends in the field of microblogging.

4 Twitter for Sentiment and Opinion Analysis

Research on the usage of Twitter for sentiment and opinion analysis only started becoming available toward the last quarter of 2009. Here, the state of the art in this particular field is highlighted.

4.1 Twitter to Gauge User Interest

Banerjee et al. [25] have presented a novel approach to discovering user interest and context by analyzing contents of tweets (in the message domain) from the Twitter Search API. Data acquisition is performed similar to other studies on Twitter mentioned above.

The user domain plays a role in this study by using user location information to isolate messages that have been composed by users in major cities [25], while user statistics allow the researchers to distinguish between active and inactive users.

The bulk of the remaining analysis comes in the message domain, using techniques to identify "tweets that capture a user's real-time interests in activities" by searching for three types of words: categories of activities (e.g. dance, food, movie, music, sports), action verbs (e.g. watch, play), and temporal nouns (e.g. today, tonight). By matching co-occurrences of such keywords in tweets, [25] managed to observe user interests and planned activities in different cities, proving that Twitter is suitable for user context analysis in terms of user interest, emotions, presence, etc. with the objective of capturing consumer data in real-time.

4.2 Twitter for Opinion Analysis: Case Study in Political Debates

An application of Twitter to analyze its users' opinions and to help annotate events is discovered by Shamma, Kennedy and Churchill [26] who investigated Twitter chatter (in the message domain) during the 2008 US Presidential Debates.

Their data collection methodology does not depend on the polling of the public timeline but rather based on the Search API (akin to [9]). By polling the API using search queries regarding the presidential debates to obtain a collection of Twitter messages, they could compare Twitter chatter to the actual debate happening in the real-world.

Their first finding corroborates literature on spike analysis, whereby they found that the "level of Twitter activity serves as a predictor of changes

in topics in the media event". They also delve in the user domain to map conversational structures between users (with respect to @user messages). This is based primarily in the message domain.

They also have shown that the matching of context can be done between the actual texts of the debate (obtained through close captioning) with the Twitter chatter. This is a form of opinion mining with regards to the collective "twittering reactions" exhibited by Twitter users toward a particular topic mentioned during the debate. As for this part of the investigation, they explore the user domain, particularly with regards to the social network graph of the friend/follower relationships between influential Twitter users.

4.3 Twitter for Marketing and Brand Sentiment Analysis

Jansen et al. [27] have proposed an automated framework for brand sentiment analysis in terms of user sentiment in relation to certain products via Twitter messages. They used the Summize service (later acquired by Twitter) to analyze "tweet sentiment" with regard to a set of nouns describing brand names. Several points in their methodology include:

- In the message domain: Jansen et al. categorized the user intention (cf. Java et al. as discussed prior [4]) into four classes: sentiment, information seeking, in-formation providing, and comments [27]. The message length of a tweet is also taken into account to study the linguistics of the user base when commenting on a brand. Co-occurrence of key terms and phrases such as personal preposi-tions are also identified in the messages.
- In the user domain: Jansen et al. tracked the volume of tweets between customers and the brand/company's Twitter account, and their frequency to determine communication patterns [27].

The findings from [27] were that certain co-occurrences of words found in tweets can correlate to the sentiments or intentions of their authors. About 19% of the total tweets "mention an organization or product brand in some way," and about 20% of these tweets also "expressed a sentiment or opinion concerning that company, product, or service". This indicates the suitability of Twitter as an avenue for mining sentiments of users.

To conclude this section, these papers summarize the vastness of information available from a microblogging service such as Twitter. The efficacy of using Twitter in traditional market research and opinion/sentiment mining serves to complement and enrich the existing findings used in the realm of conventional blogs and social media sites.

5 Human Factors on Twitter

This section surveys existing literature, both on Twitter and also in various aspects of the computer-human interaction (CHI) field, in terms of online presence, and information sharing among humans in a social setting.

5.1 User Intentions for Participating in Twitter

The intentions of using Twitter and sharing information in general have been discussed by studies investigating primarily the message domain on Twitter. In their pioneering work on Twitter, Java et al. [4] have proposed four kinds of Twitter usage intentions, i.e. "chatter, communication, information sharing, and news reporting".

Expanding on this is the work of Mischaud, who, in his Master's thesis [28], discovered that users do not only write tweets to answer the question "what are you doing?"[3] , but also for three other functions: sending messages to contacts, publishing one's thoughts, and also share "news-like information with others". This definition by [28] expands on the original list by Java by describing how Twitter is used to express ones' thoughts and describe what one is doing at a given moment; [28] also breaks down the categories of information sharing into seven distinct groups. Mischaud's work in 2007 [28] has also identified the current trend[4] of using Twitter for more than just publishing statuses about everyday minutiae.

Naaman, Boase and Lai [29] have approached this aspect from a social-awareness stream (SAS) viewpoint. They characterize message content on Twitter (as a SAS) into categories, some of which have already been discussed in [4, 28] above. Notable concepts that are unique to [29] include "self-promotion, complaints, random thoughts, [posing] questions to followers, presence maintenance [and] anecdotes". Here, they introduce the concept of using Twitter to maintain online presence and provide anecdotes to followers, which explores the motivation of Twitter users to frequently post tweets. Jansen et al.'s enumeration of user intentions from [27] included "sentiment" as one of them, which could fit under Naaman, Boase and Lai's [29] definition of "anecdotes".

[3] Twitter's original motto on their website.

[4] As of writing, Twitter has changed its motto once from "what are you doing?" to "what's happening?" to reflect the shift in its usage.

5.2 Best Practices and Typical Usage Scenarios

Several books on Twitter that focus on user participation and adoption of Twitter for marketing have also discussed about user intentions and online presence. Here, new concepts (apart from the ones in the previous paragraphs) will be given a summary.

Comm [30] wrote about the concept of "mission accomplished" tweets to inform followers of accomplishments or milestones achieved (extending the findings on online presence [29]), and picture distribution tweets (extending the concept of URL sharing by [4]). O'Reilly and Milstein [31] discussed the need for "ambient intimacy" with friends and family as a result of presence maintenance on Twitter by answering the "what are you doing?" question. Lastly, McFedries [32] and Comm [30] also highlight a current trend of Twitter usage - 'live tweeting' - which is to tweet about events live as they unfold, e.g. conferences ('conference tweeting'), trade shows and exhibitions.

5.3 Social Information Needs and Wants

Although not strictly within the realm of Twitter, Dearman, Kellar & Truong [33] have performed a field study on the information needs and sharing opportunities in normal everyday life.

Research in [33] involved volunteers manually jotting down their "daily information needs and sharing desires" in a diary which is then analyzed by the authors. They have identified 9 distinct 'information categories' and 21 subcategories in which a question or piece of shared information could be classified as.

This complements existing literature on the user intentions of microblogging and is a more detailed categorization than the grouping of 7 categories in [28]. For instance, the question "is it too cold outside to go running?" posed in [33] fits into the usage intent of "asking a question" [29] and also the 'information category' of "environmental conditions", specifically the weather.

6 Twitter in Computer-based Visualizations

The field of electronic media, computer graphics and visualizations will be given a review due to the increasing amount of attention paid to Twitter as a source of data behind several art installations and online/Web 2.0-based sites and mashups.

6.1 Twitter: Visualization and CHI studies

The process of visualization of Twitter information normally starts with the acquisition of a message timeline (e.g. using the public timeline or repeated polls of the Search API as done with other research). Studies in this regard normally take both the user and message domains into account, where information is presented in a relevant and easy to understand format, e.g. a visualization of tweets based on specified criteria, or to study a user based on his individual tweets.

First, the work of Hazlewood, Makice & Ryan [34] is reviewed. The authors have come up with a project - Twitterspace - "which is a public display of tweets published by members of [their] local community." [34]. It is a visualization whereby recent tweets by followers of a Twitter account set up for the purposes of the project are set up in a timeline-based interface to visualize the chatter of users 'belonging' to their Twitter community. This creates a 'community-at-a-glance' which aims to "[blend] the virtual space of Twitter with [their] physical community centers" [34].

Fig. 1: Hazlewood, Makice & Ryan's Twitterspace display [34].

Makice who is currently involved in research in phatics and community design using Twitter describes in a paper [35] several methods of visualization being researched at the moment, including the previously mentioned Twitterspace (Fig. 1), and an information stream 'river' metaphor.

Li [36] has developed a Twitter visualization project - Graffiter - as part of his ongoing PhD research. The basic premise behind this project is, for a given user, it creates a monthly tag cloud to visualize the most important keywords in a user's Twitter activity.

Also, as part of Li's HCI research [36], he has also introduced tags (with operator-like notation) as an extension to the normal #hashtag to allow users to explicitly express properties such as their mood, working hours, and lunch plans as part of this visualization.

6.2 Twitter Web-based Visualization Tools

Besides academic work on Twitter-based visualizations, there are several interactive Web 2.0 sites or mashups that attempt to 'make sense of' the large volume of information on Twitter. This section will detail several notable ones.

TwitterVision [37] is a mash-up between the Twitter public timeline and Google Maps to visualize Twitter updates based on the published geographic location and superimposes this onto a Google Map display to have a real-time display of Tweets based on their geographic location. The location-retrieval method behind this (and several other related websites based on the same principle) using an embedded set of location coordinates generated by GPS-enabled devices, which will then be used by Twitter in determining the user's current location.

Fig. 2: Bloch and Carter's [38] New York Times Super Bowl visualization tool, using spatio-temporal perspectives to visualize Twitter chatter.

Table 1: Summary comparison of research covered, their scope, and critique.

Research categories	Domains studied	Comments
Exploratory studies [6, 4, 3]	User: extensive friend/follower analysis, demographics, etc. Message: limited coverage	Mostly dealing with the properties of the Twitter 'ecosystem' of users such as user connections, locale, and network growth rate. Cons: message domain is studied minimally.
Message Addressivity and Forwarding on Twitter [7, 6]	User: limited to message threading between users. Message: content analysis	Deals with the existence and functions of the addressivity/retweeting syntaxes found in Twitter messages.
Twitter in Crisis and Convergence [9, 10]	User: demographic properties Message: communication styles	Deals with communication styles and user intent regarding crisis and convergence communications in a microblog setting. Pros: deals with the combined user/message domains in tandem.
Pattern Detection and User/Message Clustering on Twitter [11,12]	User: demographic properties Message: emergent statistical features	Combines both user/message domains to find out information about users. Involves mining user/message information per topic, detecting and clustering emergent features e.g. demography, habits, real-world happenings.
Viral Information Spread and Memetics [15, 16]	Not on microblogs per se.	Can be adapted for future work in tracking information spread (both user/message domains).
Trend Analysis from Existing Blogs and Social Media [22, 23, 21]	Not on microblogs per se.	Can be adapted for future work in tracking spikes and trends (primarily focusing on the message domains).
Twitter for Sentiment and Opinion Analysis [25, 26, 27]	User: social network connections, locales, habits Message: extracting opinions, and user intentions from text	Incorporation of sentiment/opinion analysis in microblogging. Pros: Twitter is a good source to harvest opinions and analysis. Field is rather new, however; as most papers are available late 2009.
User Intentions for Participating in Twitter [28, 29]	User: not studied Message: identifying cues from message text	Can be adapted for future work in justifying motivations of users from a communications/sociology /computer-supported collaborative work perspective.
Best Practices and Typical Usage Scenarios	User: not studied Message: types of messages the user base creates	Can be used in future work for applications such as opinion leader detection and spam isolation.
Social Information Needs and Wants [33]	Not on microblogs per se.	Can be used in future work for studying social information; e.g. recommender systems based on social networks.
Twitter: Visualization CHI studies [34, 35, 36]	User: social connections, habits Message: content, communications to other users	Promising field incorporating HCI and visualization studies for presentation of information to users.
Web-based Visualization Tools	User: locale, time, basic demography Message: tags, textual content and statistics	Promising field of study: requires more research and analysis in the academic context (e.g. effectiveness, potential applications, drawbacks).

Related to this is work by Bloch and Carter [38] for the New York Times, whose Flash applet (Fig. 2) visualizes Twitter activity during the Super Bowl by mapping out "location and frequency of commonly used words in Super Bowl related messages". This is not dissimilar to the use of geography and time to track the spread of a current real-life event as seen in work such as [13].

The concept of timeline visualization has also been implemented using keywords, #hashtags, and trending keywords in Twitter. This is illustrated by their use among websites such as TwitScoop [39] and various other similar sites, where interactive timelines - in conjunction with other elements such as a tag cloud and a list of related tweets - are used to illustrate the prevalence of certain popular words in current Twitter activity based on their polling of the Twitter API.

7 Comparison

The prior sections have highlighted the current state of the art in microblog research, while incorporating ideas and concepts from related disciplines that can be applied in such research. We summarize the findings of prior sections, highlighting noteworthy papers discussed, discussing any pros and cons, and also suggestions for improvements, in Table 1.

8 Conclusion

This paper has reviewed the state of the art in literature regarding Twitter as a mi-croblogging service. Incorporating ideas from blogs, email, the Web, CHI, data mining, human factors, and the like, research on Twitter is growing day by day.

As Twitter is a constantly-evolving Web 2.0 service, several changes may oc-cur; as a result, more research areas are opening up in answering previously non-existent questions, and also to improve on existing Twitter research.

This review has hopefully given the reader an insight into potential research on Twitter, specifically bridging the gap between the user/message domains on Twit-ter as the majority of existing research deals with the messages (tweets) on Twitter with little emphasis given on the underlying users behind them.

References

1. Twitter Inc.: Twitter. Available from http://www.twitter.com (2009) Accessed 8 February 2010.
2. boyd, d.: Bibliography of research on Twitter & microblogging. Available from http://-www.danah.org/researchBibs/twitter.html (December 2009) Accessed 8 February 2010.
3. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about Twitter. In: Proc. WOSN'08. (2008) 19-24
4. Java, A., Song, X., Finin, T., Tsen, B.: Why we Twitter: An analysis of a microblogging community. In: Proc. 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, Springer-Verlag (2009) 118-138
5. Huberman, B., Romero, D., Wu, F.: Social networks that matter: Twitter under the micro-scope. Technical report, Social Computing Laboratory, HP Labs (2008) Available from http://ssrn.com/abstract=1313405. Accessed 8 February 2010.
6. Honeycutt, C., Herring, S.: Beyond microblogging: Conversation and collaboration via Twitter. In: Proc. 42nd Hawaii International Conference on System Sciences. (2009) 1-10
7. boyd, d., Golder, S., Lotan, G.: Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In: Proc. HICCS-43. (2010) 1-10
8. Sutton, J., Palen, L., Shlovski, I.: Back-channels on the front lines: Emerging use of social media in the 2007 Southern California Wildfires. In: Proc. ISCRAM 2008. (2008)
9. Hughes, A., Palen, L.: Twitter adoption and use in mass convergence and emergency events. In: Proc. ISCRAM 2009. (2009)
10. Starbird, K., Palen, L., Hughes, A., Vieweg, S.: Chatter on The Red: What hazards threat reveals about the social life of microblogged information. In: Proc. CSCW 2010. (2010) 241-250
11. Cheong, M., Lee, V.: Integrating web-based intelligence retrieval and decision-making from the Twitter Trends knowledge base. In: Proc. CIKM 2009 Co-Located Workshops: SWSM 2009. (2009) 1-8
12. Cheong, M., Lee, V.: "Twittering for Earth": A study on the impact of microblogging activism on Earth Hour 2009 in Australia. In: Proc. ACIIDS 2010. (2010) (In press).
13. Cheong, M.: 'What are you Tweeting about?': A survey of Trending Topics within the Twitter community. Technical Report 2009/251, Clayton School of Information Technology, Monash University (2009)
14. Hodge, K.: It's all in the memes. The Guardian (August 10 2000)
15. Arbesman, S.: The Memespread Project: An initial analysis of the contagious nature of information in social networks. Available from http://www.arbesman.net/-memespread.pdf (2004) Accessed 8 February 2010.
16. Wasik, B.: And Then There's This: How Stories Live and Die in Viral Culture. Penguin Group (USA), New York, NY (2009)
17. Gladwell, M.: The Tipping Point: How Little Things Can Make a Big Difference. Back Bay Books, New York, NY (2002)
18. Surowiecki, J.: The Wisdom of Crowds. Abacus, London (2005)
19. Mayer, M.: What The Trend? Available from http://www.whatthetrend.com (2009) Accessed 8 February 2010.
20. Relax News: Current Twitter trends: Google Wave, 'A real wife'. Available from http://www.independent.co.uk/news/media/current-twitter-trends-google-wave-a-real-wife-1820222.html (November 13 2009) Accessed 8 February 2010.
21. Gruhl, D., Guha, R., Kumar, R., Novak, J., Tomkins, A.: The predictive power of online chatter. In: Proc. KDD 2005. (2005) 78-87
22. Choudhury, M.D., Sundaram, H., John, A., Seligmann, D.D.: Can blog communication dynamics be correlated with stock market activity? In: Proc. HYPERTEXT 2008. (2008) 55-60

23. Fukuhara, T., Murayama, T., Nishida, T.: Analyzing concerns of people using weblog articles and real world temporal data. In: Proc. WWW 2005. (2005)
24. Gruhl, D., Liben-Nowell, D., Guha, R., Tomkins, A.: Information diffusion through blogspace. In: Proc. WWW 2004. (2004) 491-501
25. Banerjee, N., Chakraborty, D., Dasgupta, K., Mittal, S., Joshi, A., Nagar, S., Rai, A., Madan, S.: User interests in social media sites: an exploration with micro-blogs. In: Proc. CIKM '09. (2009) 1823-1826
26. Shamma, D.A., Kennedy, L., Churchil, E.F.: Tweet the debates: Understanding community annotation of uncollected sources. In: Proc. ACM Multimedia 2009. (2009) 3-10
27. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Micro-blogging as online word of mouth branding. In: Proc. CHI 2009. (2009) 3859-3864
28. Mischaud, E.: Twitter: Expressions of the whole self. Master's thesis, London School of Economics and Political Science (2007)
29. Naaman, M., Boase, J., Lai, C.: Is it Really About Me? Message content in social awareness streams. In: Proc. CSCW 2010. (2010) 189-192
30. Comm, J.: Twitter power: how to dominate your market one tweet at a time. Wiley, Hoboken, NJ (2009)
31. O'Reilly, T., Milstein, S.: The Twitter Book. O'Reilly Media, Inc., Sebastopol, CA (2009)
32. McFedries, P.: Twitter: tips, tricks, and tweets. Wiley, Indianapolis, IN (2009)
33. Dearman, D., Kellar, M., Truong, K.N.: An examination of daily information needs and sharing opportunities. In: Proc. CSCW 2008. (2008) 679-688
34. Hazlewood, W., Makice, K., Ryan, W.: Twitterspace: A co-developed display using Twitter to enhance community awareness. In: Proc. Participatory Design Conference. (2008) 230-234
35. Makice, K.: Phatics and the design of community. In: Proc. CHI 2009. (2009) 3133-3136
36. Li, I.: Grafitter. Available from http://www.grafitter.com (2009) Accessed 8 February 2010.
37. Troy, D.: twittervision. Available from http://twittervision.com/ (2009) Accessed 8 February 2010.
38. Bloch, M., Carter, S.: Twitter chatter during the super bowl. The New York Times. Available from http://www.nytimes.com/interactive/2009/02/02/sports/-20090202_superbowl_twitter.html (February 2 2009) Accessed 8 February 2010.
39. Lollicode SARL: Twitscoop - stay on top of twitter! Available at http://-www.twitscoop.com/ (2009) Accessed 8 February 2010.

Part V
Software Applications

Unleash the CSS-Factor

A Social Capital Approach to the Benefits and Challenges of Corporate Social Software

Carina Heppke

Abstract 'The power of social software is undeniable in the free, anarchic world of the global internet. But what happens when you bring these tools into the constrained, policy-driven, risk-averse world of the corporate intranet where the user population is small, where expressing oneself as an individual and on a personal level can feel threatening, and where management is watching your every move' [15]. This conceptual paper takes a social capital perspective in order to explain the benefits and challenges of social software inside the firewall of organisations. Corporate social software is considered to hold great benefits for the management and the efficient use of knowledge within organisations which is regarded to become an increasingly important capability for companies in changing and challenging business environments in which adaptation, change and innovation are required to stay ahead. However, the extent to which the benefits of corporate social software will be realised by organisations depends on the way that social technologies are actually used inside the firewall. While external social technologies such as Facebook and Twitter have quickly established themselves in the daily usage patterns of a large majority of people, the usage of similar technologies within the firewall of organisations is characterised by distinct differences which will be discussed in this paper.

Carina Heppke,
University of Cambridge, Judge Business School Trumpington Street, Cambridge, CB2 1AG,
Cambridge, UK
e-mail: ch471@cam.ac.uk

1 Introduction

'Imagine a corporate world of free-flowing information from as many sources as you have employees. A world in which the information consumer controls what they consume from a menu of feeds - basing that choice on the reputation of the source, recommendations from colleagues and serendipitous discovery through social networks. Interactions are almost exclusively real-time and informal in nature [...]' [15].

During the last couple of years a (r)evolution from the web 1.0 to the web 2.0 has taken place. While 1.0 and 2.0 do not describe two distinct versions of the continuously evolving internet technology, they describe even more an evolution of technological possibilities on the one and the way that people communicate in general on the other hand. During the early years, the internet was characterised by a one-way communication. Information was presented on a website and consumed by users. With the evolution of the web towards the so-called web 2.0, the technology is increasingly facilitating not only a two-way but meanwhile multi-way communication. Social Networking Sites such as Facebook and Micro-Blogging Services such as Twitter allow pushing real-time information from one person to millions of other people in the web. Users have turned from a pure consumer role of online-content into a producer role. User generated content on websites has developed into a core element of today's participative web. The emergence of the web 2.0 is therefore characterised by two intervolving developments, firstly the evolvement of so called social technology which allows participation and multi-way communication and secondly, the adaptation of users to the possibilities that the software offers. People all over the world have quickly made sense of the evolving technology and have integrated connections on Facebook, micro blogging via Twitter and the sharing of knowledge via Wikipedia into their daily routines. Although the web 2.0 is often defined as social technology, it earns this name not only by what it offers to its users but mainly by the way that it is actually used and applied for social interaction and communication.

Because the interplay between the technology itself and the way that it is used is so vital, it raises various questions for the use of social software applications within organisations. While a 2007 Forrester survey among 119 companies still indicated resistance concerning the implementation of Corporate Social Software (CSS), Forrester also predicted a fundamental re-thinking process in the following years which would push CSS to the priority list of more and more companies. Given the rapid adaptation and success that social technologies have experienced outside companies' firewalls and the extent to which they have changed the way we communicate, interact, socialise and live our lives, the question whether the use of these technologies inside organisations might add value might appear obsolete. However, the further course of this paper will demonstrate that the value added by CSS is a complex topic with a multitude of impact factors that have to be taken into consideration.

2 The Intranet (R)Evolution

'[...] In this world, 'operational' channels dominate - glued together by an intricate web of connections between individuals, defined by individuals, bypassing the irrelevant organisational structure. How far your voice carries in this torrent of noise depends upon your reputation and what you have to say, not where you sit in any organisation chart [...]' [15].

Just like the first phase web 1.0, intranets have been and are still to a great extent characterised by a one-way communication. 'Content is produced by a small number of employees assigned to that task and all the underlying technical procedures are delegated to the IT people in charge of the security and maintenance of the system' ([5] 1).

The classic intranet 1.0 can be defined as a web-based platform for information display, where information such as the organisational structure, organisational contacts and other information is presented. Users are predominantly passive and information consumers rather than contributors of information and active participants.

CSS is in its very core still a web-based platform, however differs from the 1.0 intranet version in fundamental characteristics. CSS can consist of or be a combination of technologies that internet users are meanwhile more than familiar with such as online social communities like Facebook, wiki-technology as known from Wikipedia, micro-blogging technologies such as Twitter or video-sharing platforms such as Youtube.

Since the implementation of CSS is still in its very infancy, it is difficult to constitute *the* corporate social software. It can be assumed that companies will implement solutions which are highly personalised to the specific characteristics of the implementing company. However, in order to be able to match a CSS with company characteristics and requirements, an understanding has to be developed which risks and benefits CSS carries.

The character of intranet technology is still primarily characterised by a culture of norms, control and security, underlying the functionality of the software [9]. With the rise of the Web 2.0, people on the Internet have suddenly developed from being consumers to producers and started to extensively contribute user generated content. This radical (r)evolution of Internet usage patterns increasingly rises organisations' awareness for their intranets being rather static, rarely updated and carrying a great but unused potential regarding knowledge management [5]. CSS is considered to become of specific interest for organisations in the near future. This is based on two main reasons: Firstly, external social software such as Facebook and Twitter present the greatest phenomenon of the Web 2.0 era since they attract millions of people who connect and communicate with each other. Secondly, the user groups of these external social technologies are increasingly overlapping with organisations' employees. Thirdly, these employees use external social technologies to an increasing extent to connect and exchange information in

specific - often company related - groups. Richard Dennison, Intranet and Channel Strategy Manager at BT, puts it in a nutshell: 'When over 4,000 of your employees voluntarily join a Facebook group called 'BT', it's time to take note' [15]. This development can contain both a potential and a risk for organisations: The risk that employees make confidential information public or present the organisation in a negative way and the potential benefit of social capital development, interaction and knowledge sharing. This, as long as lying unused in the public domain, can be regarded as losing potential for competitive advantage. However, both aspects support considerations of how and to what extent these public interactions can be transferred and utilised inside the firewall.

3 E=MC^2

'[...] In this world, work truly is an activity and not a place. The boundaries between 'work' and 'play' are blurred to the point of virtual invisibility and the boundaries around organisations are permeable. The personal relationship is king[...]' [15].

The connection between CSS and an organisation's efficiency can be described with the formula E=MC^2, whereby the Mastery of each individual (human capital) multiplied by the Connections that join individuals into a community and the Communication that flows through those connections result in the Effectiveness of an organization [25, 24]. C^2 which comprises connections (structure) and communication via these connections can be defined as social capital. Social capital can be best described as a metaphor about advantage. 'Society can be viewed as a market in which people exchange a variety of goods and ideas in pursuit of their interests. Certain people or certain groups, do better in the sense of receiving higher returns for their efforts [...]. The human capital explanation of the inequality is that the people who do better are more able individuals [...]. Social Capital is the contextual complement to human capital. The social capital metaphor is that people who do better are somehow better connected' ([8] : 2).

At the core of social capital theory are networks. This automatically explains the importance of social capital for organisation studies since 'the structure of any organisation can be thought of as a network' ([26] : 26). The embeddedness and position in and within the network shapes the actions of its participants [6, 21]. In order to understand the complexity of the specific network entity in focus, its structural and relational embeddedness is of importance. Structural embeddedness refers to the density and cohesion of a network and explains how an entity is embedded in this structure [18]. Relational embeddedness focuses on the context and type of relationships within a network, i.e. the sort and strength of ties [18, 4]. 'Among the key

facets [...] are trust and trustworthiness, norms and sanctions, obligations and expectations and identity and identification' [13]. A further dimension represents 'an actor's positional embeddedness, that provides information about the information benefits of ties and networks itself' ([32] : 432).

Social capital can be both a resource and a driver to use resources in a specific way [29]. Two widely accepted definitions for social capital stem from Bourdieu and Coleman. According to them, social capital is defined as 'the actual or potential resources that stem from having a durable network of more or less institutionalised relationships of mutual acquaintance and recognition' ([2]: 2) or 'as those aspects of social structure that can be used by actors to realise their interests' ([11]: 305).

Controversial in the social capital literature is the role of network homogeneity. It is a point of major discussion whether homogeneity and therefore higher network density is more efficient in terms of knowledge creation, exchange and sharing than heterogeneity involving lower network density [28]. At the core of this debate are the 'Closure Perspective' [10, 11] and the 'Structural Hole Approach' [6, 7]. Network density or social 'closure' inside a group indicate the likely absence of 'structural holes', and are thought to foster identification with the group and a level of mutual trust, which facilitates exchange and collective action.' ([30]: 503). The 'Structural Hole Approach' in contrast describes gaps between the nodes in a network structure. Granovetter [18, 19] can be regarded as one of the main contributors to the outlined controversy. He holds that lower density enriches information exchange due to the opportunity of bridging various sub-groups of different backgrounds and stocks of knowledge.

Social capital presents a valuable perspective to challenge the potential benefits and risks of CSS. The following analysis will focus on the aspect of employee participation in CSS and as for the theoretical part on network embeddedness as a particular stream of social capital. The aspect of user participation in CSS is the most central and controversial one which is based on its dependency on various drivers [27]. The aspect of embeddedness is regarded to be crucial since it primarily illuminates the complexity underlying CSS utilisation.

4 Corporate Social Software – A Secret Weapon?

Two main factors which are assumed to impact CSS participation are trust and power. Power that is embedded in the employee's transformation from a consumer into a producer and trust in the use of CSS. CSS has the potential to drive the networking activity of an already existing offline organisational network. It fosters connectedness among employees, since it facilitates the access to information about colleagues and the connetedness among them. However, the key factor which drives the true value of CSS is the extent and

type of employee participation. This concerns primarily how employees connect in and with CSS. Do they build virtual connections with people they are already connected with in the offline corporate network or do they search for new contacts? Due to the theoretical approaches of [18, 19] and [11], the former would foster a homogenous network with strong ties while the latter drives a heterogenic network with weak connections. However, due to the embeddedness of CSS in the organisational network, the dynamics are more complex. In the context of this paper, the factor underlying the classification of homogenous and heterogeneous networks is considered to be the stock of knowledge which an individual holds. This can be driven through firstly, the private background, secondly, the professional background and thirdly, the position within a company (professional and hierarchical) of an individual. Network closure receives new dynamics in this context and blurs the concept of strong and weak ties. Scholars traditionally associate homogenous networks with rather strong and heterogeneous networks with rather weak ties. CSS however raises the question, which value e.g. the development of weak connections between two individuals with strong homogeneity has in terms of knowledge creation and sharing. When extending [18, 19] approach, the formation of connections between employees that are firstly not connected with each other in the offline-network and secondly whose stock of knowledge is most heterogeneous would provide the greatest potential for knowledge sharing and value of CSS. However, the question arises, whether the exchange between most heterogeneous groups actually provides the greatest value in terms of an organisation's efficiency. Which level of heterogeneity provides the most valuable outcome and how does the value of a stronger connectedness and exchange between rather homogenous groups (expert know-how) range in contrast to that. Furthermore, the theoretical classification of strong and weak connections in an organisational context raises the question to what extent a strong connectedness overall exists or whether a network of colleagues is most likely a network of weak ties when compared to strong connections in the private network of an employee. The characteristics of this classification are considered to have a crucial impact on the dynamics such as trust and sense of belongingness with the overall corporate network and on CSS usage.

At a general level it can be summarised that, no matter if CSS fosters the development of offline strong ties or the formation of weak ties, a stronger connectedness through CSS drives network closure of the overall corporate network to some extent with the accompanying aspects of increasing trust and identification in and with the network. However, the aspect of trust and identification presents a circular reference between the organisational network and CSS since trust is assumed to be a cornerstone for participation in CSS.

While informal information exchange happens anywhere and anytime in organisations, CSS shifts this communication to a different and simultaneously public level. The extent to which CSS drives connectedness and information exchange is therefore strongly dependent on the employees' willingness

to shift their intra-organisational communication behaviour to a more social and public level.

The aspect of trust in the context of social capital is complex. As there is regarded to be a 'widespread preference for transacting with individuals of known reputation' ([20]: 490), one can assume that the exchange of knowledge is also strongly dependent on the level and dimensions of trust manifested in the network. The embeddedness in the overall corporate context can work as a basis for trust. This sense of belongingness to an organisation as an anchor-point can create imaginary connections between employees who do not even know each other and have never met before. These imaginary ties can carry trust and the potential to drive participation in CSS. This aspect puts the role of latent ties in a completely new perspective. Latent ties are connections that are technically possible but not yet activated [22, 23]. This type of social capital has previously not received much attention. However, it is assumed that latent ties are an important driver for the creation of weak ties [22, 23] which directly uncovers again a circular reference with CSS since the software itself is likely to drive the transformation of latent ties into weak network ties by facilitating the contact between employees.

In the CSS's function as a technology, which is embedded in the existing organisational network of interpersonal connections, it has the potential for developing trust when driving connectivity and interaction between employees. In this function, CSS could even be imagined to work as a boundary object [33] when fostering the growth of bridging social ties between different sub-communities within an organisation. However, on the level of the CSS itself, its embeddedness in the organisational network is likely to work as a barrier to trust and hence to information exchange. Because communication is suddenly shifted to a public level within the organisation, power-relations, hierarchy and control come into play. These aspects lead to the question, how much social interaction can ever be possible in a CSS context and how unbiased this interaction can be. Aspects such as an organisation's culture, identity and the sense of belonging play an important role.

5 Benefits versus Challenges of CSS-Evidence from Inside the Firewall

Practical examples for differences between CSS use as opposed to internet usage patterns are manifold. A micro-blogging technology could lead to various positive effects when being implemented in an intra-organisational context. E.g. it could be used as a quick and efficient way to post a problem or question and to receive a solution. It could further be used as a status-information to inform colleagues about current unavailability. Both examples of CSS usage could lead to an increasingly efficient work process in the organisation. The former presents an innovative, creative and quick way to develop a solution

for a problem by simultaneously decreasing the likeliness of a tunnel vision because the problem is challenged in a public, intra-organisational way. The latter reduces the number of unsuccessful calls and emails (contact attempts) to a person, hence increases the efficient use of working time which could directly reflect in an increasing productivity.

British Telecom, a leading international telecommunication company, is one of the early adopters of social technologies and experiments with different types of technology inside the firewall in order to test for their value added. British Telecom currently uses various social technologies such as social networking software, wiki-technology and a video-sharing platform. The wiki technology for example is used to increase connectivity of certain divisions within the company. It presents a central location at which project related information is shared. Not only does this have the potential to increase efficiency of the work process but it also enhances innovative knowledge sharing and information exchange. The wiki-technology allows to share information received by employees at conferences, meetings and at other occasions outside the company or ideas generated in the daily work process in quasi real-time with connected employees. Particularly in highly dynamic and innovative business environments such as the one that British Telecom is active in, time can be a crucial variable in order to stay ahead in the competition. Therefore, the increase of efficiency and productivity and the real-time exchange of crucial knowledge can make a substantial difference in the competitiveness of organisations.

However, despite the outlined positive effects, the presented examples are ideal types of corporate social software usage only if the software is also used in an ideal way. Only if employees integrate CSS in their daily routines as people have done with Facebook and Twitter, will the social technology actually add value to the organisation. If 50% of employees in an organisation are highly active in updating an internal wiki whenever new information occurs, however the other 50% of employees do not access and use this information, the efficiency of CSS is questionable. In order to be of value, the use of CSS has therefore to develop into an organisational routine.

It can further be assumed that, depending on the culture of an organisation, employees might be more or less willing to post questions and problems via a micro-blogging technology to the intra-organisational community, because it might be perceived as admitting a weakness when asking the internal community for advice or help. Only in an organisational culture which fosters knowledge exchange, which accepts failure and which acknowledges that often rather weird questions have to be asked for great ideas to develop, will the full benefit of a corporate social software flourish. Google might be an excellent example for an organisational culture in which the potential benefits of CSS can be fully used. In 2007, Google's back then product chief (Matt Glotzbach) stated in an interview that it is ok to fail wisely. "At Google, we really focus on failing wisely. [...] There is no penalty for failure. In fact we

encourage it because if you are not failing it means you are probably not trying." [1].

However, while CSS has to become part of the daily routines of employees to add value to an organization, exactly this constitutes a potential challenge. The more employees are involved in CSS usage, the greater is the chance that processes become inefficient, that waste is produced and that the quantity of information enhances inefficiency. In this sense, the logic of $E=MC^2$ has a tipping point at which a greater connectedness and communication via CSS does not lead to an increasing efficiency for the organisation but can even turn into the opposite. A popular argument originates from Dunbar, who claims 150 to be the 'magic number' for group sizes, whose exceedance will automatically result in a fragmentation of the group (Gladwell, 2000). The more individuals are connected with each other, the more each community-member trusts the whole network and therefore uses it for the exchange of information, hence, the more the group as a whole benefits. In this sense, it could be argued, that an increasing quantity of interconnections through CSS usage with concurrent constant quality would lead to an overall rise of social capital. However, going back to the level of the individual CSS user, it can be assumed that the strength of the interconnectivity of users and therefore the value of these connections would suffer from a growth beyond a specific tipping point. The amount of time invested in the CSS as a whole spreads among more ties and therefore loses quality and intensity. The more information is exchanged via CSS and the more knowledge shared, the more does a CSS raise completely new challenges for knowledge management and efficient work processes, because the connections and information resulting from CSS usage start to require a management themselves.

6 Conclusion

'A key lesson is to focus on the value social media tools can deliver rather than the risks. If you dwell too much on the risks, you'll never leave the starting gates. There are risks, but the potential benefits are huge' [15].

The previous argumentation has revealed that the implementation of CSS means an intrusion in the complex dynamics of an intra-organisational network. The value of CSS is strongly grounded in the level and extent of user participation. This leads to the question whether a critical mass is required. E.g. in terms of knowledge management the question arises if 'the wisdom of crowds works with a smaller population of people, or will the result be unbalanced opinion rather than truth?' ([15]: 1) and whether the use of social software in a corporate context follows the same principles than external Social Networking Sites such as Facebook. Yet, even the dynamics of Facebook, although it is significantly more established than CSS, are only understood

to a small extent. It exists a broad agreement that once a critical tipping point is reached, viral effects push network effects forward. However, there is no understanding why and when a critical mass is reached and what makes users leave or decrease their activity at some point in time.

The analysis of this paper has provided evidence that the dynamics of CSS might differ significantly from external social technologies. Udell [31] for example expects network effects to be much slimmer in a corporate context. Given the absence of direct competition within a company, this would strengthen the importance of trust and benefits as important drivers of participation. Overall, a question for further investigation is to what extent organisations can support or even enforce participation though incentives, internal marketing or the use of power and manage participation..

This paper has demonstrated that the reservations of companies concerning the implementation of CSS are not unjustified. On the one hand, looking at CSS from a social capital perspective has clearly outlined a significant potential. On the other hand, great uncertainties have been uncovered in terms of participation and the unpredictability concerning the development of the overall employee network of an organisation. Metaphorically speaking, CSS provides intra-organisational communication with legs and feet but until now there exists no real understanding how far, how fast and whether at all this communication will start to walk, nor if acclaimed benefits of the Web 2.0 such as the 'wisdom of crowds', hold in a corporate context. Further research in this field can provide a deeper understanding of how the complex impact factors underlying CSS success are all connected. These results are regarded to offer important insights for organisations in order to balance potential benefits against risks of CSS implementation. Of particular importance is the tipping point at which the value added by CSS might be not positively correlated with the number of users and the extent of CSS usage but shift to the negative. CSS enhanced connectedness and information exchange and the extent to which both have to be managed in order to become usable and increase organisational efficiency form and input-output equation which offers a breeding ground for future investigation. An understanding of this equation is assumed to have a strong impact for developers of CSS as well as companies which plan to implement social technologies. The argumentation of this paper suggests that firm size and hierarchical and organisational structure might be important impact variables on the extent to which CSS correlates with an organisation's efficiency. An understanding of the principles and dynamics that different types of CSS could have on an organisation given its specific characteristics will be valuable when it comes for organisations to decide which type of CSS will add the greatest value.

References

1. Boulton, C. (2007). Google Product Chief: Its Okay to Fail Wisely. http://www.eweek.com/c/a/Messaging-and-Collaboration/Google-Product-Chief-Its-Okay-to-Fail-Wisely/. Accessed 12/02/2010
2. Bourdieu, P. (1980). Le capital social: Notes provisoires. Actes de la Recherche en Sciences Sociales 31, 2-3.
3. Bourdieu, P. (1986). The forms of capital. In: J.E. Richardson (1986). Handbook of Theory for Research in the Sociology of Education, Westport: Greenwood Press.
4. Brown, R. (1965). Social Psychology. New York: Free Press.
5. Buffa, M., Sander, P., Grattarola, J.-C. (2004). Distant cooperative software development for research and education: three years of experience, In proceedings of CALIE'04, Grenoble, France.
6. Burt, R. S. (1982). Towards a Structural Theory? Academic Press, New York.
7. Burt, R.S. (1992). Structural Holes: The Social Structure of Competition. Cambridge, MA: Harvard University Press.
8. Burt, R.S. (1997). The contingent value of social capital. Admin. Sci. Quart. 42, 339-364.
9. Ciborra, C. (2000). A Critical Review of the Literature on the Management of Corporate Information Infrastructure. In Ciborra el al. (eds.) From Control to Drift, Oxford University Press, 15-40.
10. Coleman, J. S. (1988). Social capital in the creation of human capital. Atrzer. J. Sociology 94, 95-120
11. Coleman, J.S. (1990). Foundations of Social Theory. Cambridge, MA: Harvard University Press.
12. Coser, R. (1975). The Complexity of Roles as Seedbed of Individual Autonomy. In Coser, L. (1975). The Idea of Social Structure: Essays in Honor of Robert Merton. New York: Harcourt Brace Jovanovich.
13. Cummings, S., Heeks, R., Huysman, M. (2003). Knowledge and Learning in Online Communities in Development: A Social Capital Perspective. Working Paper Series, IDPM, University of Manchester
14. Fontana, J. (2007). Unified communication's twists and turns ensure slow arrival on corporate networks. NetworkWorld.com, http://www.networkworld.com/news/2007/101607-microsoft-unified-communications-analysis.html. Accessed 16 October 2007
15. Dennison, R. (2007a). BT Web 2.0 adoption case study. Inside out blog by Richard Dennison, BT, http://richarddennison.wordpress.com/bt-web-20-adoption-case-study/. Accessed 25 March 2008
16. Dennison, R. (2007b). The future of internal communications...?. Inside out blog by Richard Dennison, BT http://richarddennison.wordpress.com/my-articles/the-future-of-internal-communications/. Accessed 25 March 2008
17. Forrester (2007). Global Enterprise Web 2.0 Market Forecast: 2007 To 2013. Forrester Research
18. Granovetter, M.S. (1973). The Strength of Weak Ties. American Journal of Sociology 1973, 78(6), 1360-1380
19. Granovetter, M.S. (1983). The Strength of Weak Ties: A Network Theory revisited. Sociological Theory, 1 (1983), 201-233.
20. Granovetter, M.S. (1985). Economic action and social structure: the problem of embeddedness. The American Journal of Sociology [AJS], 91(3), 481 - 510.
21. Granovetter, M. (1992): Problems of explanation in economic sociology, In Nohria, N. and Eccles, R. (eds), Networks and Organizations: Structure, Form and Action (Boston, MA: Harvard Business School Press)
22. Haythornthwaite, C. (2002). Strong, weak, and latent ties and the impact of new media. The Information Society, 18(5), 385-401.

23. Haythornthwaite, C. (2005). Social networks and Internet connectivity effects. Information, Communication & Society, 8(2), 125-147.
24. Krebs, V.E. (2008). Managing the Connected Organization. http://www.orgnet.com/MCO.html. Accessed 28 March 2008
25. Leana, C.R. & van Buren, H.J. (1999). Organizational Social Capital and Employment Practices. The Academy of Management Review, 24(3) (Jul., 1999), 538-555
26. Nohria, N. & Eccles, R.G. (1992). Face-to-Face: Making Network Organizations Work. In N. Nohria & R.G. Eccles (eds.). Networks and Organizations: Structure, Form and Action. Boston, MA: Harvard Business School Press, 288-308
27. Patmore, J. (2007). Silicon Valley comes to Cambridge, Conference Talk, University of Cambridge, Judge Business School, 11/12/2007
28. Portes, A. (1998). Social capital: Its origins and applications in modern sociology. Ann. Rev. Sociology 24, 1-24.
29. Putnam, R. D. (2000). Bowling Alone: The collapse and revival of American community. New York: Simon & Schuster.
30. Reagans, R. & Zuckermann, E.W. (2001). Networks, Diversity, and Productivity: The social Capital of Corporate R&D Teams. Organization Science, 12(4), July-August 2001
31. Udell, J. (2006). Reinventing the intranet. InfoWorld. Accessed 19 March 2008. http://www.infoworld.com/article/06/04/05/77011_15OPstrategic_1.html.
32. Van Wijk, R., van den Bosch, F.A.J., Volberda, H.w. (2003). Knowledge and Networks. In Easterby-Smith, M. & Lyles, M.A. (2003). The Blackwell Handbook of Organizational Learning and Knowledge Management, Malden: Blackwell Publishing
33. Wenger, E. (1998). Communities of Practice: Learning, Meaning, and Identity. Cambridge, UK/Cambridge University Press.

Extending SQL to Support Privacy Policies

Kambiz Ghazinour, Sampson Pun, Maryam Majedi, Amir H. Chinaei, and Ken Barker

Abstract Increasing concerns over Internet applications that violate user privacy by exploiting (back-end) database vulnerabilities must be addressed to protect both customer privacy and to ensure corporate strategic assets remain trustworthy. This chapter describes an extension onto database catalogues and Structured Query Language (SQL) for supporting privacy in Internet applications, such as in social networks, e-health, e-government, etc. The idea is to introduce new predicates to SQL commands to capture common privacy requirements, such as purpose, visibility, generalization, and retention for both mandatory and discretionary access control policies. The contribution is that corporations, when creating the underlying databases, will be able to define what their mandatory privacy policies are with which all application users have to comply. Furthermore, each application user, when providing their own data, will be able to define their own privacy policies with which other users have to comply. The extension is supported with underlying catalogues and algorithms. The experiments demonstrate a very reasonable overhead for the extension. The result is a low-cost mechanism to create new systems that are privacy aware and also to transform legacy databases to their privacy-preserving equivalents. Although the examples are from social networks, one can apply the results to data security and user privacy of other enterprises as well.

Kambiz Ghazinour, Sampson Pun, Maryam Majedi and Ken Barker ,
University of Calgary,
Department of Computer Science, Calgary, AB, Canada
e-mail: {kghazino,szypun,mmajedi,kbarker}@ucalgary.ca

Amir H. Chinaei,
University of Puerto Rico (Mayagüez campus), Department of Electrical and Computer Engineering, Mayaguez, PR, USA.
e-mail: ahchinaei@ece.uprm.edu

1 Introduction

The Structured Query Language (SQL) is at the core of millions of database applications, many of which were implemented during the past decade. This chapter introduces an important extension onto SQL, which gives increased expressive power to its security model so it can support today's privacy requirements. The extension is simple to understand, easy to implement, and inexpensive to maintain.

Before proceeding to next sections, this work is motivated with running examples from the area of online social networks. (Social networks are good examples of applications that require user-centricity and are very vulnerable to privacy violations.) Even though current social networks, such as Facebook, Flickr and My-Space, publish a privacy policy indicating the usage of personal information within their network, they do not provide a system-level or user-centric protection to enforce the policies. Facebook and Flickr have chosen to join a third party auditing firm to resolve disputes once they occur, and MySpace relies on its internal policy to enforce such rules. Currently, other social networks do not provide more effective privacy enforcement either. This chapter proposes a user-centric enforcement of privacy policies applicable to a variety of web applications, such as social networks, e-health, e-government, etc. The proposal extends the nearly ubiquitous SQL language but the core ideas are readily extendible to other modern database languages, such as XQuery. The balance of this section illustrates the extension's requirements.

1.1 Requirements for Extension

Similar to Rosenthal and Sciore's contributions [10], three criteria during the extension are considered: modularity, compatibility, and simplicity. Modularity allows system vendors to provide a straightforward implementation of their systems to support the additional privacy provided by this increment. Furthermore, it helps system administrators and end-users to gradually adopt the new features at their discretion. Backwards compatibility is critical when transforming legacy systems to their privacy-preserving equivalents. It also provides an easier implementation because it is built on existing technology. Simplicity, which is partially ensured by modularity and compatibility, is a key characteristic for the extension to be accepted by the market. These three criteria together make the advances simple to understand, easy to implement, and inexpensive to maintain.

1.2 User Privacy Requirements

The key elements of data privacy are purpose, visibility, generalization, and retention [3, 5, 6]. Purpose specifies the legitimate reasons to access a specific piece of data or information. Purpose-based access control models require a set of finer grain operations. For example, in standard SQL, the SELECT statement is a coarse operation; but, in purpose-based systems, a SELECT operation with a "survey" purpose may be distinguished from the one with a "marketing" purpose. Purpose-based access control systems must provide a mechanism to enforce the intended purposes within the operations. Visibility specifies the legitimate users who can access particular data, for a legitimate purpose. Visibility constrains the set of users who can access data with respect to an operation and a purpose. For example, the visibility of operation "write on a wall" could be defined as group "friend" in a social network. Thus, not every user can write a message on that wall.

Generalization specifies the level of anonymization required when presenting a given data value in response to a legitimate access request from a user for a particular purpose. The level of generalization, denoted by n, depends on the data type, its domain, and the amount of k-anonymity and l-diversity defined by applications [8, 11]. For instance, for an age attribute, when $n=1$, this may mean that age can be disclosed if it is generalized to be in a range between $\pm$ 5; and for a postal code attribute, n=1, may mean that the postal code can be disclosed if it is generalized by hiding its right most digit.

Retention specifies an expiry condition (based on time, period, number of accesses, etc.), after which the data is no longer accessible, even for legitimate users, purposes, and/or appropriate level of generalization. Furthermore, each application may define user-defined privacy constraints (UDC) to capture additional privacy semantics. UDCs can be exploited to implement constraints such as obligations, legitimate time and location of access, or legitimate number of accesses.

1.3 Organization

The rest of this chapter is organized as follows. Section 2 extends the current SQL security model towards supporting the privacy policies. Section 3 addresses the underlying privacy catalogues. Section 4 highlights the implementation issues. Section 5 provides a comparison among other proposals. Finally, Section 6 concludes this chapter and touches on the future directions.

2 Extending SQL to Support Privacy Policies

Current social networks, in particular, offer a very limited degree of privacy protection to their users. This is true for most Internet applications, in general. Most privacy protection is implemented at the application level, allowing adversaries to exploit (back-end) database vulnerabilities. This exposes users to an increased possibility of privacy violations. Many Internet applications are unique in that they require both mandatory access control and discretionary access control mechanisms. For MAC models, it is compulsory to consider privacy policies in an earlier stage. It is proposed to enforce them when data containers (for instance, tables in a relational model) are being formed. Section 2.1 reviews the SQL's CREATE TABLE statement; and, Section 2.2 extends the statement to support corporate-based privacy policies. In order to support the DAC requirements, the GRANT and REVOKE commands are enhanced. Section 2.3 through 2.7 reviews these two SQL security commands and extends them to support privacy policies.

2.1 Overview of CREATE TABLE

In SQL, one can create tables by using the CREATE TABLE statement. It is assumed, in a MAC system, only security officers (e.g. database administrators in relational models) are privileged to exercise this statement. Following is a simplified syntax of the CREATE TABLE statement [1, 2, 3]:

```
CREATE TABLE tbl
(col dataType [column_options]...,[table_options]...);
```

in which *tbl* is the name of table being created; *col* is the name of a column included in table *tbl*; *dataType* indicates the type of data stored in column *col*; (examples of data types are INT, CHAR, DATE, etc.;) column_options allow different rules to be applied to column *col*; (examples of these rules are forcing the column to have a certain default value, and ensuring the column is not NULL;); there could be more than one column in a table; table_options are rules which can be applied to *tbl* (examples of these rules are limiting the minimum and maximum number of rows in *tbl* or setting a password for table access).

2.2 Extended CREATE TABLE

In this section, CREATE TABLE is extended to include two optional parameters: one for generalization function and one for retention, as follows.

```
CREATE TABLE tbl
(col dataType [column_options] [transformFunction]...,
[table_options]..., [retention]);
```

in which *tbl*, *col*, *dataType*, column-options, table-options, and partition-options are as described in Section 2.1; moreover, transformFunction is an optional parameter for column *col*; any column storing sensitive information should choose a generalization function; this function is used to generalize the data of the column to a certain level, specified as a parameter that is defined by privacy policies (cf. Section 3.1); retention is an optional parameter that specifies the expiry condition of each row of table *tbl* and, as described in Section 1.2, there could be a variety of conditions for implementing retention. The default values of the *transformFunction* and retention are "none" and "all", respectively. This means, if the corresponding clause not used, no anonymization is applied (generalization is "none") and a granted privilege never expires (retention is "all"), respectively. The following example illustrates the extended CREATE TABLE further.

Example 1: The following statement creates a table that stores address of members of a social network. Also, it specifies that when a particular member address is going to be accessed, function *generalizeAddress* generalizes the address before the disclosure. (Again, note that the amount of generalization is defined by privacy policies.) Also, since the retention clause is absent in the statement, the data never expires.

```
CREATE TABLE MembersAddress
(ID INT, Address VARCHAR(50) generalizeAddress);
```

2.3 Overview of GRANT

The SQL GRANT statement allows users to assign a subset of their privileges to other users (within authorized IDs). A particular user cannot access an object unless the appropriate privilege is held. The set of privileges applicable to a particular object depends on the object type. Examples of object types are functions, tables, and views. The focus is on the privileges applicable to tables (and views) using a simplified syntax of such GRANT statements [7]:

```
GRANT op (col) ON tbl TO u [WITH GRANT OPTION];
```

where *op* is a comma-separated list of privileges (operations). Examples of operations are SELECT, UPDATE, DELETE, etc.; *tbl* is either a base table or a view to which op applies; *col* is a comma list of columns of *tbl*; u is the authorization ID of the user (or the user group) to whom the privilege is granted; and finally, the optional WITH GRANT OPTION clause, may be used to allow u to further grant op to others.

Example 2a: The following statement allows Alice to update columns Name and Email of table MyProfileView, which could automatically be defined-for every user-as a view on a base table that includes all profiles.

```
GRANT UPDATE (Name, Email) ON MyProfileView TO Alice;
```

Example 2b: The following statement allows Alice to query any column of table MyProfileView.

```
GRANT SELECT ON MyProfileView TO Alice;
```

2.4 Modified GRANT

The GRANT statement is extended to include an optional FOR clause. The FOR clause specifies a privacy policy that includes restrictions such as purpose, generalization, retention, and user defined constraints (udc). Moreover, an optional TO clause is introduced to support visibility in the WITH GRANT OPTION. The extended syntax of GRANT is as follows:

```
GRANT op (col) ON tbl TO u
[WITH GRANT OPTION [TO visibility]]
[[FOR purpose, generalization, retention [, udc] ]];
```

where *op, col, tbl,* and u are as described in Section 2.3. (1) visibility specifies a comma-separated list of users (authorization IDs) to which user u is allowed to further delegate operation *op* on table *tbl*; (2) purpose specifies the only intention of user u for which u is allowed to apply operation *op* on table *tbl*; (3) generalization specifies the level of anonymization that must be applied to the disclosing part of table *tbl* just before disclosure; and (4) retention specifies an expiry condition by which the granted privilege is no longer available. Finally, udc is an optional parameter that allows for user-defined constraints. As a syntactic sugar, the extension also allows the user to specify several privacy policies by including several FOR clauses in one GRANT statement. Note that there are two application-dependent features in the extension: udc and generalization. For example, a member of a social network may want to be notified via email whenever their profile is being visited. As an example of application dependency of generalization, a member may prefer to show their address to some other members in a generalized form, such as "district", "city", or "country"-rather than specific or micro data. Finally, to retrofit legacy systems' privacy and to be flexible with new systems, all ingredients of the extension are optional in the syntax and the database catalogues are designed to comply with this feature too. The default values of visibility, purpose, generalization, and retention are "all", "any", "none", and "all", respectively. This means, if the corresponding clause is not used, a granted privilege can be further granted to any user (visibility is "all"), a granted privilege can be used for any intention (purpose is "any"), no anonymization

is applied (generalization is "none"), and a granted privilege never expires (retention is "all"). The rest of this section provides two examples to illustrate the modified GRANT.

Example 3a: Assume Bob executes the following GRANT:

```
GRANT SELECT ON MyProfileView TO Alice
WITH GRANT OPTION TO Friends
FOR "sendgift",0,"20101231", Inform(email);
```

This means 1) Alice is privileged to query Bob's profile if her intention is to send a gift and the above date has not passed; moreover, Alice is allowed to further grant the same privilege (based on the policy stated in this statement) to every user who is a member of group Friends. 2) The disclosing data will not be generalized (generalization=0). 3) Alice must also inform Bob, via email, of every query she sends. Note that in this example, the udc clause has been utilized to implement obligation. Also, note that -for readability- scalar parameters are used for purpose, generalization, and retention, but one can use functions to allow more expressivity.

Example 3b: The following statement grants two privileges to Alice: one is a privilege to delete rows of table Event and the other is to update attributes Telephone and Address of that same table; but, she can exercise these privileges only when her intention is to change an address (before 20101231) or for bookkeeping (before 20100720). Data will not be generalized in any case, and if Alice exercises DELETE or UPDATE to change an address, she must inform the data owner by email.

```
GRANT DELETE, UPDATE(Tel,Adrs) ON Event TO Alice
FOR "change address", 0, "20101231",Inform(email)
FOR "bookkeeping", 0, "20100720";
```

2.5 Overview of REVOKE

The SQL REVOKE statement allows a user to reverse any previous GRANT statements, which results in taking back the established privilege. The following simplified REVOKE statement syntax revokes a privilege on a table from a specific user [7]:

```
REVOKE [GRANT OPTION FOR] op ON tbl FROM u;
```

where *op* is a comma-separated list of privileges (operations); *tbl* is either a base table or a view to which *op* applies; u is the authorization ID of the user (or the user group) from whom the privilege is to be revoked; GRANT OPTION FOR, if used, specifies that only the grant option of the privilege

is revoked but not the privilege itself.

Example 4a: The following statement revokes Alice's privilege of selecting from table MyProfileView.

```
REVOKE SELECT ON MyProfileView FROM Alice;
```

Note: In contrast to GRANT, the user cannot apply REVOKE on some columns of a table. It is applied instead to all columns of the table.

Example 4b: The following statement revokes Alice's privilege of updating any column (including columns given in Example 2a) of table MyProfileView.

```
REVOKE UPDATE ON MyProfileView FROM Alice;
```

2.6 Modified REVOKE

REVOKE is extended to include an optional FOR clause. The following depicts the extended syntax of REVOKE:

```
REVOKE [GRANT OPTION FOR] op ON tbl FROM u
[FOR purpose];
```

where *op*, *tbl*, and u are as described in Section 2.5. Purpose specifies a comma-separated list of intentions for which user u is no longer privileged to access *op* (or its grant option) on table *tbl*. Similar to the extended GRANT statement, the FOR clause is optional. The default value of purpose is "all", which means-if the FOR clause is not used, user u can no longer use privilege *op* regardless of intentions. The balance of this section provides two examples to illustrate the extended REVOKE further.

Example 5a: Bob can cancel the privilege he granted to Alice in Example 3a by executing the following statement:

```
REVOKE SELECT ON MyProfileView FROM Alice
FOR "sendinggift";
```

or, by executing the following statement:

```
REVOKE SELECT ON MyProfileView FROM Alice;
```

Note that, in the latter case, the SELECT privilege is revoked from Alice for any purpose.

Example 5b: The following statement revokes the privilege from Alice to delete a record from Event only when her purpose is to cancel party.

```
REVOKE DELETE ON Event FROM Alice FOR "cancel party";
```

3 Model Semantics

This section describes the semantics of the model operationally, using privacy catalogues and the relational model. Section 3.1 introduces the underlying privacy catalogues. Section 3.2 addresses how the model changes the SQL data manipulation language; in particular, a new algorithm for the SELECT statement is illustrated.

3.1 Privacy Catalogues

To represent the semantics of the model, two privacy catalogues are proposed, namely *SysPrivacyPolicies* and *SysTransforms*. These catalogues represent user privacy policies and generalization functions, respectively. These would ideally be implemented as part of the system catalogues. However, for simplicity, and to demonstrate the approach's utility transparently, they have been implemented as a set of standalone tables. This allows for legacy systems to utilize the extension without the need for a major upgrade to the database engine itself.

(a) SysPrivacyPolicies

gtee	gtor	table	col	op	purpose	g	ret	udc	vis
Alice	Bob	MyProfileView	*	SELECT	sendgift	0	101231	Inform (email)	Friends

(b) SysTransforms

Table	column	genFuction
MyProfileView	Address	generalizeAddr
MyProfileView	Picture	generalizePic
MyProfileView	Age	generalizeAge

Fig. 1: Privacy catalogues

Fig 1 illustrates the privacy catalogues. *SysPrivacyPolicies* represents a user privacy policy consisting of ten components, namely gtee, gtor, table, *col*, *op*, purpose, g, ret, udc, and vis. Values stored in gtee and gtor specify the grantee and grantor, respectively. The values of table and *col* specify the attributes of a base table on which the policy is defined. The value stored in *op* specifies the privilege that the grantee can exercise on the attribute but only if there exists a legitimate intention (stored in purpose). Furthermore, g (generalization) values are used to specify the level of anonymization that should be applied to the attribute before the disclosure through the query

processor (it is used as an input parameter in generalization function). The ret value specifies the expiry condition of the whole policy. The udc (user defined constraints) is used to capture application-dependent conditions, such as obligations. Finally, the vis attribute specifies the user groups that are visible to the *op* for further grants. Fig 1(a) illustrates the instance of *SysPrivacyPolicies* after executing Example 3a (described in Section 2.4). It specifies that Alice can grant the SELECT privilege on all columns of table MyProfileView to every member of Friends, conformant to other constraints, such as purpose, generalization, retention and any user-defined constraints. Note that it conforms to SQL's rule that the policy details specified for a privilege are applied to its grant option too. Both mandatory and discretionary access control models (MAC and DAC, respectively) are supported with these system catalogues. For MAC privacy requirements, users will populate the *SysPrivacyPolicies* catalogue in accordance with the privacy policies users agreed upon. DAC privacy requirements will be interpreted and updated using Grant and Revoke statements described in Section 2.4 and 2.6, respectively.

SysTransforms represents the corresponding privacy transformation functions of each column of each table. It consists of three attributes: table, column, and genFunction. Attribute genFunction represents the corresponding generalization function of a table column. This catalogue is populated automatically when a base table is created, if the optional generalization clause is used. Fig 1(b) illustrates three transform functions applicable to different columns of table MyProfileView. For example, the first row specifies that when a particular value of column Address of table MyProfileView is to be accessed by a user, function *generalizeAddress* is applied to that value. The amount of generalization, the input to the function, is varied for different users and is specified in SysPrivacyPolicies.

Finally, the retention condition is maintained in a hidden column of the base table. Then, every time a row is being accessed, its corresponding retention is checked to verify whether or not the row is accessible.

3.2 Extended Data Manipulation Language

This section describes how the SQL data manipulation language (DML) can be extended to support the modified SQL statements. Recall that the proposed enrichment concepts for the SQL security model are purpose, visibility, generalization, retention, and user defined constraints. As described in Section 3.1, generalization can be automatically applied when a column is being accessed. Similarly, retention can be checked when a column is being accessed: if, due to the retention, the column is no longer accessible for the corresponding operation and purpose, the operation will be rejected. User defined constraints (e.g. inform(email)) can be enforced as a function call in the transaction executing the operation. If the function returns false, the

transaction is aborted. It is important to note that each operation is treated as one database transaction in which all components (such as purpose and obligation) of the operation will either commit or abort. In other words, the transaction manager is exploited to enforce the purpose of usage and the user-defined constraints such as obligations.

To support purpose, one must extend all DML operations by adding a new clause, namely REASON, by which the users (or processes) specify what their purpose is in requesting to execute the operation. If the corresponding grantee, operation, and purpose are found in the privacy catalogues (cf. Fig 1), the operation may continue conformant to the corresponding generalization, retention, and user defined requirements. For the rest of this section, the SELECT statement is considered to further illustrate this approach. Note that SELECT is used to retrieve information from a database and poses the greatest threat to user privacy. (Interested readers can readily apply the extension to all other relevant DML operations.) The SELECT operation is extended as follows:

```
SELECT      column
FROM         table
[WHERE     condition]
[REASON     purpose];
```

Note that the extended clause is optional to support backward compatibility. However, if privacy catalogues are available then the user who queries the database must have a REASON for it, and the reason must match the corresponding tuple in the privacy catalogues or the query is rejected.

Example 6: Recall Example 3a, in which Bob granted the SELECT privilege to Alice. Now, Alice can query the database as follows:

```
SELECT Tel FROM MyProfileView REASON "send gift";
```

Note that the query will be rejected if the retention date (20101231) has passed. Furthermore, the transaction manager requires Alice to inform Bob by email before the result is disclosed. Finally, note that if Alice submits a query with a purpose (e.g. bookkeeping), which is not already granted to her, the query is rejected.

Fig 2 illustrates an algorithm to interpret the extended SELECT statement in this model. In Line 1, the authorization ID of the user and all of the user's member groups are identified. In Line 2, the result is initialized to empty. In Line 3, the SysPrivacyPolicies is queried to verify if the data requester (denoted as gtee) is permitted to execute the operation (which is SELECT in this case) on the table with respect to the specified purpose. If so, in line 4, those rows of the result that meet the retention condition are generalized (using corresponding generalization functions stored in SysTransforms) and then disclosed. Otherwise, an empty set is disclosed.

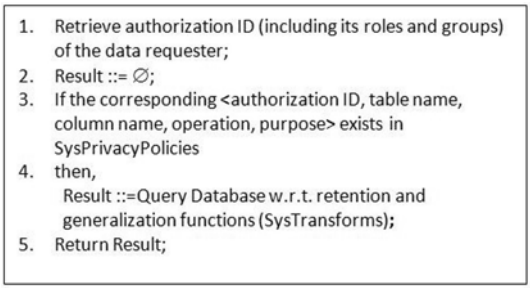


Fig. 2: Extended SELECT algorithm

4 Complexity Analysis

Fig 2's algorithm is no more complex than the conventional ones, where the privacy rules are checked in the core systems, so the incremental impact on system performance is negligible. In particular, only Lines 3 and 4 in Fig 2 introduce new overhead to the conventional algorithms. Line 3 imposes a search cost of O(r log r) where r is the number of privacy rules in the catalogue. Line 4 imposes an additional predicate check (for retention) and a generalization function, both of which have a complexity of O(1). Note that the complexity of Line 3 is already being paid in the conventional approach because only more rules are introduced in support of user preferences. Thus, the incremental complexity is zero but the size of r has increased slightly. To provide this support for user preferences in the system catalogues themselves would still require this price be paid. Furthermore, an experiment to study the cost of the extension is performed using a data and query set generated from the TPC-H benchmarking standard.

In the first experiment, the size of base table is kept as 10K records and the size of privacy catalogues is increased to study its effect on the privacy overhead timings. Fig 3 illustrates the results of this experiment. As the size of catalogue is increased to 20K the corresponding times for each MAC and DAC model increased only 30.3 milliseconds and the total privacy overhead time that contains both MAC and DAC increased by 61.3 milliseconds. During this experiment, since the size of base tables did not change the query execution time remained constant. It is also noticeable that since the approach used for privacy of MAC and DAC were close the timings of them are also nearly equal.

In another experiment, the size of privacy catalogues is kept constant to 20K records and the sizes of base tables is increased to study and compare the time spent for privacy checking and the time to query execution. Fig 4 illustrates that as the size of base tables are increased to 50K the query execution time increases to 532.18 milliseconds. Hence, the time of privacy checking, 61.3 milliseconds, is negligible.

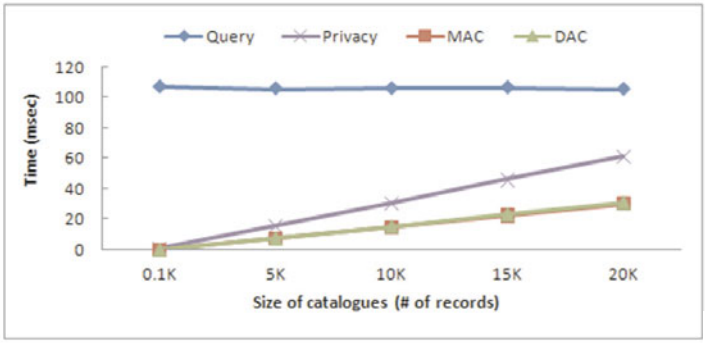


Fig. 3: Comparing size of base tables and corresponding privacy times

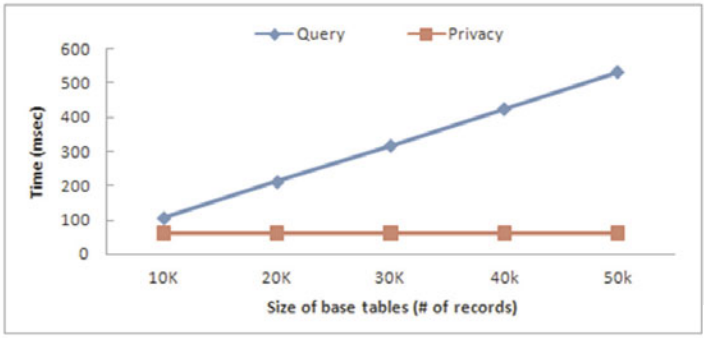


Fig. 4: Comparing size of base tables and corresponding privacy times

Fig 5, shows a broader view of the last two experiments in which the size of privacy catalogues and base tables are increased. It is demonstrated that as the size of these datasets increases, the privacy catalogue has quite a small growth than the query execution time.

Fig 6 illustrates the experimental results from another perspective. As the size of base tables increases, the overhead of privacy checking becomes more trivial.

4.1 Implementation

A prototype has been also developed using MS Visual Basic.NET to investigate the extended data manipulation language in practice. The prototype acts as a wrapper application that mediates between the users and the underlying database management system. It consists of a text-base interface,

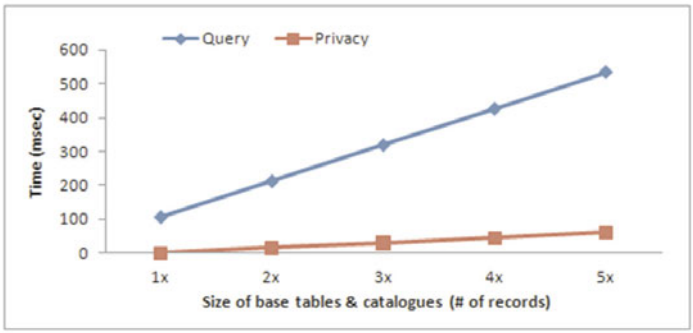


Fig. 5: Comparing size of base tables and catalogues and corresponding query and privacy times

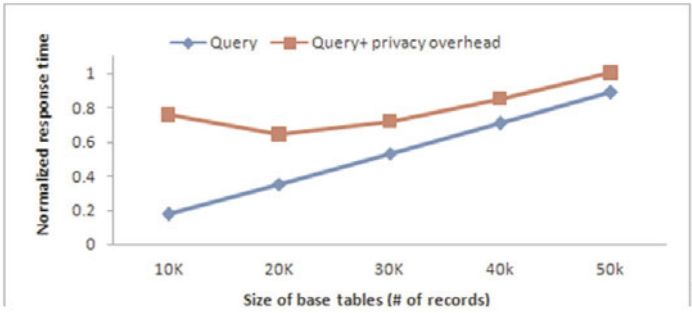


Fig. 6: The effect of privacy overhead as the size of base tables increases

a query parser, and a privacy control module. The parser accepts only basic SQL syntaxes, most of which were explained in previous sections of this chapter. The privacy control module, then, updates the privacy catalogues (illustrated in Fig 1) and passes the query to the underlying DBMS, which is MySQL 6. Fig 7 abstracts the communication flow among these components.

In particular, users type SQL statements in the interface of the query parser. The query parser separates extended SQL privacy syntax and MY/SQL 6.0 SQL syntax. The query parser then provides the privacy control module with the table and columns the user is interested in as well as any privacy values provided (such as the reason for access). The privacy control module is responsible for two different areas, enforcing a privacy-based access control and altering the results to the generalization stated in SysTransforms. To enforce the privacy-based access control, it will verify whether the user has a valid reason for performing the requested operation on the table and column. If this privacy tuple exists within SysPrivacyPolicies, the query is forwarded to the DBMS; otherwise, an empty set is returned to the user. Moreover,

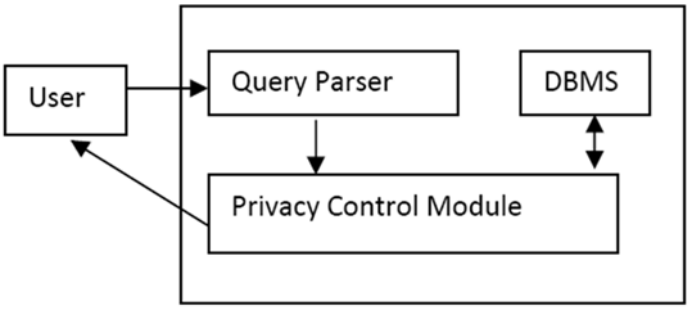


Fig. 7: Communication flow

for SELECT statements, the module alters the resulting tuples based on the generalization function (stored in SysTransforms) and the level of generalization (stored in SysPrivacyPolicies), and only releases those tuples that meet the retention condition.

5 Related Work

Rosenthal and Sciore have proposed to extend the SQL security model by adding a new predicate to the GRANT statement [10]. They consider simple predicates (e.g. working hours) mainly related to environment conditions. They also discuss the proper revocation time and the frequency of evaluating the grant conditions, but a syntax or semantics is not developed.

Barker and Rosenthal demonstrate how policies, specified in stratified logic, may be exploited to help security administrators recognize the behavior of access control systems [4]. The policies are transformed into a subset of SQL statements to guarantee that only legitimate users have access to the data. Agrawal et al. propose a model based on GRANT statements to describe a language construct capable of specifying restrictions at the row, column, and cell levels [1, 2]. In their approach, GRANT statements can be translated to security views. They also introduce a new algorithm for translating P3P privacy policies [5] into their proposed construct.

Staden and Olivier also acknowledge that users must declare their intention to access a data value and their intention must be compared against their profile [12]. The model differs from this proposal in that they only consider purpose for the SELECT statement. Moreover, they state that privacy preservation cannot be accomplished by using a discretionary access control model only. Instead, they propose a Hybrid-DAC model for access control in which the system security officer binds purposes to data and then grants a proper privilege to the user to access that data. In the work presented in this

chapter, the data owner can define their own policy and assign privileges to other users with legitimate purposes. Staden and Olivier [13] subsequently complete their previous proposal by extending the REVOKE statement too; whereas, this work addresses both GRANT and REVOKE simultaneously and in an integrated way.

6 Conclusion

In this chapter, the SQL security model is extended to support privacy policies in a social network. The policies support well-known privacy concepts such as purpose, visibility, generalization, and retention. Furthermore, the extension is simple, modular, and backward compatible so it is applicable to new designs or legacy systems. The model is supported with semantics defined operationally in a relational model. In particular, relevant details affecting the underlying catalogues and their supporting algorithms are discussed.

Obviously, end-users will be completely unaware of the details of syntactic extensions that support privacy, however it is critical that appropriate mechanisms be developed to transparently capture this information from the users. This is a key element to implementing this increased privacy functionality so it is essential that work be undertaken on these kinds of human factor issues. Another direction that is intriguing academically is to determine how to extend the idea to FLWOR expressions, the workhorse of XQuery language, which are applicable to semi-structured and object repositories so these techniques can be applied effectively to environments that are not driven specifically by relational database engines. Success in this area will have a substantial impact on providing better privacy in social networks such as Facebook. Finally, this privacy-centric approach, has developed a comprehensive framework that addresses both enterprise and user privacy concerns, simultaneously.

References

1. R. Agrawal, J. Kiernan, R. Srikant and Y. Xu. "Hippocratic Databases". Proceeding of the 28th International Conference. on Very Large Databases (VLDB 2002), Hong Kong, China, 2002. pp. 143-154.
2. R. Agrawal, P. M. Bird, T. W. A. Grandison, G. G. Kiernan, S. I. Logan, and W. Rjaibi, "Extending Relational Database Systems to Automatically Enforce Privacy Policies", Proceeding of 21st ICDE, Japan, 2005. pp. 1013-1023.
3. K. Barker, M. Askari, M. Banerjee, K. Ghazinour, B. Mackas, M. Majedi, S. Pun and A. Williams, "A Data Privacy Taxonomy", Proceeding of BNCOD09, England, 2009. pp. 42-54.

4. S. Barker and A. Rosenthal, "Flexible security policies in SQL", Proceeding of fifteenth Annual IFIP Working Conference on Database and Application Security, Canada, 2001, pp. 167 - 180.
5. L. F. Cranor, Web Privacy with P3P. O'Reilly Media, 2002.
6. S. Finestone, "Privacy: Where do we draw the line?", Public Works and Government Services, Canada, 1997.
7. Grant privilege statement, ANSI/ISO/IEC International Standard (IS). Database Language SQL, Part 2: Foundation (SQL/Foundation). 1999. P. 588
8. A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity", Proceeding of ICDE, USA, 2006, pp. 24-35.
9. S. Pun, A. H. Chinaei, and K. Barker, "Twins (1): Extending SQL to Support Corporation Privacy Policies in Social Networks, Proceeding of Advances in Social Networks Analysis and Mining ,Greece, 2009.
10. A. Rosenthal and E. Sciore, "Extending SQL's grant and revoke operations, to limit and reactivate privileges", IFIP Workshop on Database Security, The Netherlands, 2000, pp. 209-220.
11. L. Sweeney, "k-anonymity: A model for protecting privacy", International Journal of Uncertainty Fuzziness and Knowledge Based Systems, 2002, pp. 557-570.
12. W. J. C. van Staden, and M. S. Olivier, "Extending SQL to allow the active usage of purposes.", Lecture Notes in Computer Science, Volume 4083, Springer, 2006, pp. 123-131.
13. W. J. C. van Staden and M. S. Olivier, "SQL's revoke with a view on privacy", Proceeding of SAICSIT, South Africa, 2007, pp. 181-188.

nCompass Service Oriented Architecture for Tacit Collaboration Services

David Schroh, Neil Bozowsky, Mike Savigny, and William Wright

Abstract nCompass is a flexible, Service Oriented Architecture (SOA) designed to support the research and deployment of advanced tacit collaboration technology services for analysts. nCompass allows a significantly larger number of individual analytic capabilities, applications and services to be integrated together quickly and effectively. Service integration results are described from several computational tacit collaboration experiments conducted with open source intelligence analysts working with open source data. Key to nCompass is the technical framework and unique analytic event logging schema that supports context sharing across diverse applications and services. It is by combining the analyst with shared context across multiple advanced computational capabilities in a system of systems that a breakthrough in collaborative open source analysis can be achieved. This paper introduces the nCompass framework and integration platform, describes key nCompass core services, and provides results on functional synergies achieved through technology service integration with nCompass.

1 Introduction

In response to - and with feedback from - researchers investigating tacit collaboration computational services for Open Source Intelligence (OSINT) analysts, we have designed and developed a flexible, component-based Service Oriented Architecture (SOA) for the goal of supporting the research of synergistic advanced technology services for analysts, and the subsequent deployment of these advanced capabilities. This framework - its design, specifications and implementations - is called nCompass.

David Schroh, Neil Bozowsky, Mike Savigny and William Wright ,
Oculus Info Inc.,
e-mail: {dschroh,nbozowsky,msavigny,bwright}@oculusinfo.com

Key to nCompass is the technical framework and analytic event logging schema to support context sharing across applications and services, including user modeling, information and expert recommendation, computational linguistics, reasoning and image processing services. It is by combining the analyst with shared context across multiple advanced capabilities from a wide variety of third parties in a system of systems that a true breakthrough can be achieved. In the next section we present our objectives in designing and implementing the nCompass Service Oriented Architecture. In Section 3 we describe technologies that were leveraged to achieve these objectives. Section 4 provides a tacit collaboration scenario to illustrate the use and impact of nCompass integration and new analytic event services. Section 5 introduces nCompass as both an architectural framework and a reference implementation, with descriptions in Section 6 of core services necessary to support our approach to tacit collaboration. In Section 7, we review experimental OSINT analysis uses of nCompass, and resulting impacts on service integration, experiment design, and tacit collaboration through context-sharing. Section 8 briefly discusses related work, and in Section 9 we conclude and suggest future work.

2 Objectives

To enable quick and effective integration of multiple individual computational analytic capabilities, we set out, with iterative feedback from other research teams, to design and implement the nCompass SOA framework and integration platform. The objective of the nCompass framework was to create a single unified environment into which third party analytic components could be easily integrated to produce new computational collaborative services. The integrated system would provide applications and services for tacit collaboration, suitable for OSINT challenges in exploiting massive unstructured data that is common to the domain.

The integration framework needed to provide access to functions and data without overhead costs, allowing efficient transfer of information among components, as well as customization of capability by specific end-user organization and by analyst. The framework also needed to support advanced visualization and analyst activity capture. Analytic event logging was implemented to support context sharing across applications and services. Every third party component contributed via logging to the analyst context. Every third party component had access to the common pool of activity logging. This enabled improved analytic collaboration and information sharing among Web 2.0 applications and services that were particularly suitable for OSINT analysis.

Within the widely distributed and diverse organizations, information can be difficult to discover or access. Analysts "don't know what they don't know" [1]. It is also difficult to discover other analysts with expertise and insights

relevant to the task at hand. A key goal of nCompass in providing a framework to enable improved information awareness, was to not impose on the analyst any additional procedural or cognitive strain. The approach to achieving this goal was through tacit collaboration. In contrast to explicit collaboration, tacit collaboration allows analysts to discover important information and relevant or complementary expertise which they are unaware of, based on their actions and the actions of other analysts and computational services that identify similar interests and similar information.

3 Technical Foundations

3.1 Oculus nSpace

To define requirements for the nCompass framework, we leveraged functional integration experience with the Oculus nSpace system of systems for OSINT analysis [2]. nSpace is a visual analytics work environment for unstructured data with novel information triage, evidence marshalling and sense-making capabilities. It is implemented with an Ajax browser front end with multi-tier computational and data services. nSpace also serves as a functional integration platform that marshals a variety of third party tools and data sources for analysts. Using open web services interfaces and protocols, nSpace integrates computational resources such as reasoning services, agent-based modeling and advanced computational linguistic functions, including entity extraction, supervised and unsupervised clustering, and automatic ontology construction. As a test bed for new technologies, nSpace is a platform enabling analysis science. The impact of new technologies deployed into the nSpace platform can be seen through features such as side-by-side comparison of results from alternative tools, an integrated workspace to perform experiments in a whole analytic workflow context, and a workspace to move seamlessly between components without loss in data or task context.

nSpace uses open web services interfaces and protocols [3] for components to exchange data about information objects critical for the process of analysis (e.g. hypotheses, evidence, models). This experience with functional integration led the way for the nCompass approach.

3.2 Service Oriented Architecture (SOA)

In a Service Oriented Architecture (SOA), the capabilities within an application are exposed as services [4] [5]. Each service is autonomous, reusable, stateless and discoverable. Capabilities within an application that are suit-

able to be services correspond to a strong business activity or recognizable business function. Services should be coarse-grained, and reusable, and suitable for well-defined interfaces. This allows existing capabilities to be easily recomposed into new applications to solve unanticipated problems.

Components in a SOA should be loosely coupled, highly interoperable, and have platform- and development technology-independent access mechanisms. Web service standards define ways for applications to compose themselves of coarse-grained, reusable components with well-defined interfaces. Standards for message packaging are often coupled with machine-processable interface descriptions such as Web Service Description Language (WSDL) to make these services more easily consumable.

These characteristics led to the adaptation of a Service Oriented Architecture as the basis for the nCompass framework. For the enterprise, standardizing on a SOA would allow the leveraging of existing capabilities within the organization. Enterprise IT would also be able to draw on new research, knowing it can be integrated quickly. For R&D, a SOA would allow the quick exploration and assessment of combinations of new computational capabilities. A SOA approach would also provide the R&D community a roadmap for transition from research into production. The use of a standards compliant SOA platform would give developers a single way of packaging capabilities for multiple customers and solutions. It was our belief that one-time investment in the engineering required to work in a SOA would reduce the cost for researchers to collaborate and explore multiple technology synergies on an ongoing basis. As discussed in Section 7, experiments sponsored by the IARPA Incisive Analysis program [6] demonstrated that nCompass allows a significantly larger number of individual analytic capabilities, applications and services to be integrated together quickly and effectively.

3.3 Tacit Collaboration Approach

To improve information awareness through tacit collaboration, the nCompass SOA framework needs to support three core technical capabilities. Analysis Modeling Services use captured indicators of analytic activity to build models of the user's analytic context and discern the user's information needs. Information Modeling Services build models of information objects and their relationships to each other. And finally, context-aware services use these models of the information space, and user models, to enhance the analyst's information awareness [7].

Previously, the Pacific Northwest National Lab (PNNL) developed the Glass-Box environment for instrumentation of analyst workstations to log user activity [8]. This software was designed to capture low level detail, such as mouse clicks and keystrokes. Creating robust heuristics to infer high level analytic activity from these low level events was difficult and error prone.

A resulting design objective for nCompass, therefore, was to incorporate a new framework for capturing higher-level, more meaningful indicators of analytic activity to be made available to user modeling services. The nCompass SOA framework has been integral in connecting computational services that improve information awareness through tacit collaboration services. Document recommendation services use models of analytic context to search for information that corresponds to the user's information needs. Adaptive information retrieval services provide re-ranking of search results, presenting high-value information to the analyst based on an evolving user model. User modeling services match models of user context to one another to make recommendations of relevant or complementary expertise.

Web 2.0 technologies that leverage social networking and crowdsourcing are directly applicable to OSINT analysis. Social bookmarking [9], information recommendation engines [7] and web mining applications [2] are some of the OSINT tacit collaboration technologies that have been loosely coupled through the collection of indicators of analytic activity and interest by the nCompass SOA platform.

4 Scenario Illustrating Use and Impact

The following scenario demonstrates how four OSINT analysts, separated both in time and space, are able to share work in progress, discover important new information, and discover each other through tacit collaboration. This scenario was used in an nCompass integration experiment. All names are fictitious.

- Emily Baker is a senior OSINT analyst and an expert in nuclear proliferation. Today she is investigating possible transfers of nuclear technology from CountryA to CountryB. She has created a network of key players in her nSpace2 Sandbox and now digs deeper by issuing a query in TRIST to retrieve documents about "proliferation" and CountryA's "PersonX."
- User modeling capabilities have been subscribing to her analytic activities - searches conducted, documents exploited, markups and annotations applied - via the Analysis Log Service (ALS). When she issues her query, the integrated system recognizes she is pursuing a new line of inquiry and automatically searches the repository for documents she hasn't yet seen.
- A document is recommended that refers to CountryC's Prime Minister. Emily learns that his son was involved in transferring dual-use technology to another country of interest through a corporation connected to an associate of PersonX.
- Emily updates her nSpace Sandbox with this new information, planning to next investigate whether there is also a connection to CountryB, but she is interrupted by an urgent request for a different tasking. Rather than

suspend this current investigation, she places a note in her Sandbox and shares it with Adam Andersen, a junior analyst in her office.

- Adam sees what Emily has been thinking about a possible connection between CountryA and CountryB via CountryC. He queries for more information involving proliferation, PersonX, CountryB and CountryC, and finds a document that appears to confirm Emily's suspicions. He tags the document with keywords that aid him in organizing information of value.
- While Adam has been working, his activities have been logged and his user modeling service is automatically updated. The system now recommends another analyst who has been working on related tasks. Adam decides to follow up on the recommendation by contacting Gabriel Martinez, an imagery exploitation specialist at another organization. The system has alerted Adam that Gabriel has been investigating smuggling of dual-use technology into CountryB, and Adam wants to learn more.
- At yet a third organization, Jason Risdal, a Mideast Affairs specialist, is working on determining which countries are supplying CountryB with components for uranium enrichment. His query results in a collection of retrieved documents and recommended analysts.
- At the top of the list is Emily Baker. Jason clicks on her name to access her Analyst Profile page, where he finds references to information with which she has recently been working. He's interested in the document with the CountryC connection, as well as the set of tags associated with it.
- Jason clicks on the tag "smuggling," and sees that Gabriel Martinez has tagged more documents with this keyword than any other analyst. Clicking through to Gabriel's Analyst Profile page, Jason learns of Gabriel's extensive experience investigating smuggling of weapons and weapons components. Jason contacts Gabriel, who may be able to assist in finding visual evidence of weapons components smuggling into CountryB.
- Jason has now been connected to Gabriel through the tag "smuggling," which he found associated with a document in Emily's collection, even though it was Adam, not Emily, who had applied the tag. Adam's routine action, done for his own benefit, provided the means for connecting Jason to Gabriel. Tacit collaboration is the key to enabling efficient insights such as this.

5 nCompass

nCompass is designed as a flexible, component-based SOA framework which supports the research of synergistic advanced technology for analysts, integrating reasoning services [10], agent-based modeling [11] and advanced computational linguistic functions [12] [13] including entity extraction, supervised and unsupervised clustering, and automatic ontology construction

[14], and the deployment of these advanced capabilities. nCompass is an open SOA platform based on open web standards: [15]

- Messaging
  - SOAP 1.1
  - HTTP 1.1
  - WS-Addressing 1.0
  - MTOM 1.0
- Service Description
  - XML 1.0
  - XML Schema
  - WSDL
- Service Publication and Discovery
  - UDDI
- Security
  - HTTP Over TLS
  - TLS 3.0
  - SSL 3.0
  - X.509

nCompass is compatible with other enterprise SOA frameworks. The service interfaces are implementation agnostic, enabling deployment into existing enterprise architecture.

nCompass fulfills several roles and its design reflects these motivations. It is a research platform for managing the complexity of groups of services, allowing researchers who are combining efforts to spend more of their resources doing research. It is an experimental platform for collaborating with other researchers, exploring functional synthesis, prototyping concepts and reviewing with analysts. nCompass is also a deployment platform which includes a robust, tested reference implementation of key infrastructure services.

During deployment of new analytical services, nCompass is able to leverage enterprise SOA environments, integrating with existing enterprise SOA architecture and resources. The nCompass reference implementation also provides default capabilities, using open source SOA products that can be used as place-holders until specific enterprise products are in place. Default capabilities can also be used to minimize impacts on production systems until specific enterprise capacities are established.

6 nCompass Core Services

6.1 Analysis Log Service (ALS)

The Analysis Log Service (ALS) is a web service that collects and provides a repository for records of Analysis Log Events (ALEs). These logged events are captured high-level indicators of analytic activity and are used by Analysis Modeling Services to build models of individual and collective analytic context. These high level indicators are the foundation to this new approach for tacit collaboration services. The models of analytic context facilitate modes of tacit collaboration between analysts, and shared analytic context across applications that may not otherwise be integrated.
The design and usage of the ALS is based around a SOA infrastructure:

- Users interact with Applications, typically delivered in a web browser.
- Business Services deliver capabilities to Applications as web services.
- Business Services and User Modeling Services interact through open SOA specifications.

As users interact with Applications, these services report significant user events in the form of ALEs sent to the ALS, as shown in Figure 1. Business Services neither report nor directly consume ALEs, but obtain all information about the context of analysis by querying Analysis Modeling Services.

User Modeling Services obtain information about analysis activities and behaviors primarily by issuing requests to the ALS. They issue requests on whatever schedule and frequency most appropriate to the particular dimension(s) of analytic context they endeavor to model. User Modeling Services then supply their models to context-aware Business Services. User Modeling Services consume ALEs from the ALS, but do not (typically) supply ALEs back to the ALS [16].

6.1.1 Analysis Log Events (ALEs)

The ALS supports an open specification in which each Analysis Log Event consists of a Message schema, and a single Event element which contains one or more Object elements.

The Event element is defined in a taxonomy of Event classes such as Search, Assess and Retain. In specifying these classes, the overarching goal is to define the smallest set of classes such that records of events of a given class are consumed by at least one Analysis Modeling Service, or records of events of a given class are required for metrics assessment and evaluation purposes. The objective is to provide sufficient precision in differentiating between events for analysis modeling services to produce effective models of analytic context, without rendering the taxonomy intractable and too difficult for model

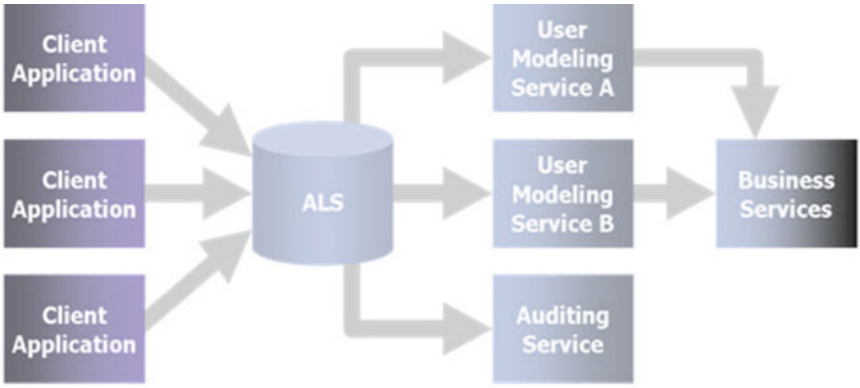


Fig. 1: Analysis Log Service (ALS) architecture.

producers to use. Each event class is extendable, so that event producers can generate events that do not fit exactly into the taxonomy, or that can contain data that is not part of the standard event definition, without being bound by the specification governance process. Each Analysis Log Event message contains Objects of various kinds. The Object class taxonomy defines Entity types along with their internal structures. There are a small number of base classes for different kinds of Objects, and some auxiliary types or classes. Entity is the base object that Resources, Relations and other Objects all inherit from. Resources include a wide variety of objects that might be transferred, excerpted, annotated, etc., including all sorts of text documents, images, audio files, and video data. All Object types have an extensions element that can contain service or application specific XML.

Analysis Log Service design and usage is predicated on the availability of a persistent store of information objects accessible across the enterprise. Information referenced in ALEs cannot be trapped within individual applications, but instead must be externally available for use by other services and applications.

6.2 Content Management Service (CMS)

The Content Management Service (CMS) is another core nCompass service that provides access to a persistent data store required to support the Analysis Log Service. It consists of a standard Simple Object Access Protocol (SOAP) web service interface to COTS and GOTS content and document management systems and can be used to store information objects ranging from individual arguments by a particular user, to large document collections. Applications can store their information objects in the CMS, either

natively or for archival purposes. The CMS interface allows for retrieval of those artifacts, and discovery through keyword search.

The CMS SOAP interface supports standard operations such as Store and Retrieve. Content can be stored into the CMS by providing document metadata, as well as the document content. Each object stored in the CMS is assigned a unique identifier, which is returned when the object is stored. It is this CMS ID that is used to retrieve the content, as well as share it with others. The CMS supports the SOAP Message Transmission Optimization Mechanism (MTOM) specification [17] for retrieving content as binary attachments, rather than encoded text. This greatly increases the performance of using a web service to transmit and store binary data.

Metadata can be associated with an object upon storage, such as author, title, or Uniform Resource Locator (URL). Arbitrary fields are also supported. A metadata search service interface allows for searching on metadata. Optional additional modules are also supported, including a keyword search service and utilities for importing, exporting, indexing and processing the contents of the CMS.

Indexing content for keyword search is an optional feature in the CMS. Content added to the CMS's data store can be flagged as being suitable for keyword indexing. Logical collections can be configured to use the internal keyword indexer, based on the open-source Apache Lucene indexer, or can be configured to use an external, third-party indexer, such as a Google Search Appliance [18] or IBM OmniFind Yahoo Edition [19]. Logical collections that represent enterprise content services that support their own indexing will pass search requests directly to those services.

The CMS also supports a Representational State Transfer (REST) interface, both for retrieval and for discovery. For retrieval, all objects in the CMS can be retrieved via a Universal Resource Indicator (URI) for discovery. The CMS supports the OpenSearch 1.1 specification for information retrieval.

A plug-in architecture allows a single CMS instance to support multiple virtual content collections. Each named virtual collection can be stored in the same physical repository, or they can be located across multiple vendor implementations, all accessible through a single web service interface. This same plug-in architecture allows multiple implementations of content stores to be accessible through the same CMS interface. Implementations of the CMS interface support open source content stores including the XML content store eXist, and the relational database Apache Derby, as well as adapters for commercial products such as Oracle and MarkLogic.

For the researcher, or small-scale production environment, the CMS provides a standard way to store documents, data, metadata, annotations and other content in a repository that is searchable and from which every object can be retrieved as a URI. For the enterprise, the CMS can provide a data store that scales up. The adapter framework allows existing enterprise content repositories to be made available through the CMS interface so that

existing capabilities that use the CMS can leverage enterprise resources with little or no additional modification.

Both the Content Management Service and the Analysis Log Service depend upon the availability of a unified authentication scheme. This is required to provide a common database of users, trusted communication between services, and to enable the controlled flow of information across application boundaries.

6.3 Authentication Management Service (AMS)

The Authentication Management Service (AMS) provides a common web service for managing credentials and user attributes across applications and services. The service presents a standard workflow for web applications that require authentication. By using a common authentication system, applications and services share a common database of users, and analysts do not have to log into multiple different systems. A verified user identity is also essential in providing accurate user modeling and tacit collaboration services.

The authentication service is not just for analysts; it is also for services. Just as users must authenticate with an application, so must that application authenticate with services that it consumes. This allows for trusted communication between services, and protection of information transmitted.

Using standard SOAP and REST access methods, users can be authenticated by either username and password against the web service, or through a client-side X.509 digital certificate. Once authenticated, applications and services identify the user by an authentication token. In the case of password-based sessions, this token is generated by the AMS. For TLS and SSL server configurations that support client-side digital certificates, this token is based on the Distinguished Name (DN) found in the certificate.

This token is passed along to other applications and services to represent the authenticated session. Downstream services can check the token, to ensure it is valid. The token can expire, or be revoked, ending the user's session. It is this standard authentication token that allows for single sign-on between applications.

The AMS also provides a user management interface for viewing common user attributes shared across applications. Behind the AMS web service can exist any proprietary authentication store; the reference implementation allows for a text file, relational database or LDAP directory. By implementing the service interfaces that are part of the specification, other authentication stores can be used.

6.4 Group Management Service (GMS)

The Group Management Service (GMS) provides a common web service to define groups for access control and social networking. The goal of the GMS is to bring the definitions of groups of users, and their roles, outside the boundaries of individual applications and services. Externalizing group definition from individual applications allows for management of group- and role-based access control. It also enables end users to create and manage their own Communities of Interest (COI) that cut across application boundaries. Externalization of groups, such as the standard "friends" list, enables a broad variety of new social software to be brought to bear on analytic challenges.

The GMS consists of a SOAP interface that defines methods for creating groups, adding users to groups, defining user roles within groups, building hierarchies of groups, and setting access control lists (ACL) for management of these groups. These access control lists allow for the creation of public, private, and semiprivate groups. A group can be set public, so that anyone can join. Alternatively, a group can be semiprivate, requiring an invitation to get in, which anyone already in the group can provide. Finally a group can be private, and only the moderator of the group can add new members.

The GMS allows for manipulating the ACL of each group, for the purposes of configuring COIs. Mandatory access control on information objects, such as due to sensitivity or privacy concerns, can be performed by applications based on group definitions. If information objects are stored in the nCompass Content Management Service (CMS), access control can be configured there.

The GMS also supports the Group and People portions of the REST API from the OpenSocial specification. This specification defines a common API for social applications. There are many websites implementing OpenSocial, including Engage.com, Friendster, hi5, Hyves, imeem, LinkedIn, MySpace, Ning, Oracle, orkut, Plaxo, Salesforce.com, Six Apart, Tianji, Viadeo, and XING [20]. The GMS supports the retrieval of an analyst's "friends" list through this API, allowing existing applications that support OpenSocial to quickly support the GMS. The GMS extends the OpenSocial REST interface to support other group management operations.

Together, the Authentication Management Service and Group Management Service enable users and services to authenticate with security across multiple applications. This is a necessary capability to support authorized access for a distributed network of users, information objects and tacit collaboration services.

7 Experiments and Results

The following describes integration experiences and results with three experiments that demonstrate the nCompass impact on system-of-system integration efficiencies for OSINT analysis work environments.

7.1 Ease of Integration

Over the course of a two year period from 2006 to 2008, we participated as one of twelve research groups combining individual capabilities with the goal of producing an increase in analytic capabilities through tacit collaboration via a combined system of systems. nCompass served as the integration platform through the course of three integration experiments. The experiments were designed to investigate and demonstrate collaboration computational services to enhance analyst and system effectiveness, including information sharing through tacit collaboration, enhanced information value of items reviewed by analysts, and increased analyst effectiveness.

The first integration experiment demonstrated SOA integration of all 12 participating research capabilities focused on information sharing through tacit collaboration. The goal of the first experiment was to evaluate how well research capabilities, in a broad spectrum of maturity from concept through to ready-to-deploy, could be integrated in a Service Oriented Architecture, with a resultant benefit for the analyst. The first nCompass platform was deployed, consisting of a service bus and message framework. All of the research teams offered up their capabilities through either request-reply or event-driven web services.

The second experiment demonstrated increased functional complexity. Teams of researchers worked across their individual project boundary to leverage concepts from other researchers. Common service offerings were established in the nCompass platform, including the Analysis Log Service, Content Management Service, and Authentication Management Service. Functional integration focused on analysis logging for "analyst-aware" applications, data finding users instead of users finding data, social networking for tacit collaboration in analytic communities, and analysis auditing and review.

The third experiment consisted of an end-to-end solution using analyst modeling to enable expert recommendations, document and data recommendations, cross-tool and multi-modal shared analytic context, and enhanced information retrieval in an OSINT analysis scenario.

A key result confirmed by design specifications and project scheduling was that, while the second and third experiments each involved increasingly complex integrated systems of systems, the timeframes for planning, design, integration, and quality assurance grew progressively shorter from ten weeks to three weeks. The nCompass SOA proved critical in providing the capability

to investigate, efficiently, breakthrough "combinatorics" of different research capabilities.

7.2 Impact on Experiment Design

In the fall of 2007, we supported the National Institute of Standards and Technology (NIST) in conducting an experiment focused on evaluating the software system known as Hydra 3 to demonstrate several key principles for enhancing analyst and system effectiveness, and set a baseline for further studies. Hydra 3 was an integrated system, that did not use nCompass, but combined the information triage and flexible analytic workspaces of Oculus nSpace, the entity extraction and document categorization capabilities [13] of the Fair Isaac Text Analysis Engine (TAE), and the natural language search and ontology generation [14] of Lymba's Power Answer and Concept Explorer (PACE).

The experiment consisted of two groups of users: a treatment group using Hydra 3, and a baseline group using Google search and Microsoft Word. Each group consisted of eight OSINT analysts who searched the corpus of approximately 30,000 documents harvested from the Internet, varying in type from news articles and digests, to blogs and forums. The task was an assessment of Country X's political leadership influences, and the analysts were asked to use the Analysis of Competing Hypotheses (ACH) methodology [21] to produce a report using a specified, detailed Microsoft Word form. ACH is a tool for weighing alternative hypotheses, helping an analyst to minimize cognitive limitations and bias. Instruments included report ranking, logging, scaled and open ended questionnaires.

Experimental results on the impact of the Hydra 3 system on analytic effectiveness were inconclusive with differences in analyst reports difficult to assess. However, in terms of workload, analysts in the baseline group were observed to perform more queries rendering more results to scan through, while the treatment group significantly reduced the amount of work required to find and save information relevant to the task.

The PNNL GlassBox environment was used to instrument the analyst workstations, recording detailed information on usage of the technology. The logged GlassBox data served to verify proper logging of Analysis Log Events by the integrated Hydra 3 system, and to validate the use of ALE data for measuring characteristics of analytic activity.

The effectiveness of the experimental process in this earlier non-nCompass experiment was compared against that of other later experiments in which nCompass served as the integration platform. This earlier experiment did not take advantage of the nCompass SOA to facilitate open integration, and instead relied on multiple custom integration points. A key conclusion derived from the comparison of this experiment to others was that more time was

spent on engineering and system testing, and less on experimental design and the research goals, when the experiment could not leverage the nCompass SOA [22].

7.3 Tacit Collaboration Through Context-sharing

In the summer of 2008, as part of a focus on evaluating tools that provide user modeling capabilities, SET Corporation tested their User Modeling Service (UMS) [23] in the context of an nCompass integrated multi-component system. The experiment was developed to test three key areas of functionality: [24]

- Virtual Interest Group recommendations (data was captured online, and experimental recommendations produced off-line)
- Adaptive Information Retrieval (re-ranking of Google document results, with online user judgments)
- Document Recommendations (system generated queries based on the user's model, with online user judgments)

We generated a new version of the nCompass ALS to support sense-making analytic events in this experiment, and hosted the ALS, CMS and AMS for the experiment. We also provided two additional components. First the analysts used the nSpace2 web-based OSINT analysis environment as their workspace [25]. In this workspace they issued queries, organized information, and viewed documents. The queries issued in the nSpace2 environment were used during the Augmented Information Retrieval evaluation stage. Information organized in the nSpace2 Sandbox was used to establish a baseline for document recommendations. Events in querying, reading, and information organization were all used to build the user models.

A re-ranking web service collected search results on behalf of the user, submitted them for re-ranking, and populated rich web forms with the re-ranked documents to the analyst for judgment. For Augmented Information Retrieval, this service collected the last query issued by the user in nSpace2, submitted it to Google and collected 50 search results. These search results were passed to the re-ranking system to be ordered based on the user model. These reordered results were then mixed with the original Google results, and presented to the user in a rich form, again for user judgment of the effectiveness of the re-ranking.

For Document Recommendations, the service collected a system generated query based on the user model, and submitted it back to Google. The search results returned were similarly submitted for reordering. The re-ranked results were mixed with a baseline generated at the start of the user's session, and presented to the user for judgments.

All searches, results, reordering, mixing, baselines, treatments, log files, and documents retrieved were recorded and archived during the experiment. The archives were made available to researchers for analysis.

With nCompass SOA as the underlying integration platform, researchers were quickly able to combine capabilities to effectively demonstrate breakthroughs from shared context across a combined system of systems. Integrating several existing and new analytic services into a powerful Web 2.0 application platform in the nCompass SOA environment saved significant integration time, compared to similar previous experiments, allowing researchers to focus on strong experimental design. Reported results of the experiment analysis indicated "significant impact of using user models to enhance finding better information in documents and in finding other people to work with" [26].

8 Related Work

Context can be broadly defined as "any information that can be used to characterize the situation of an entity" [27]. Much work in the area of context-aware computing has focused on awareness of a situation in physical environments, with elaboration of associated data schemas [28] and technical frameworks [29].

With the Analysis Log Event schema, and the technical framework of the nCompass SOA and Analysis Log Service, we are endeavoring to provide shared analytic context.

Enhancing active collaboration sessions has been proposed through sharing tool state, and finding indicators of significant analytic events through examining textual messages between collaborating analysts [30]. In contrast, our goal is to enable tacit collaboration, without requiring explicit intent or intervention on the part of the user.

Some work has examined implicit human computer interaction, using physical sensors to provide context and alleviate the need for explicit input by the user [31]. Again, this work focuses on awareness of situation in a physical environment.

Our approach to enabling tacit collaboration relies upon unobtrusively capturing indicators of analytic activity as the user interacts with an instrumented workspace.

Prior to the work described in this paper, the Pacific Northwest National Lab (PNNL) developed the GlassBox environment for instrumentation of analyst workstations to log user activity [8]. This software was designed to capture low level detail, such as mouse clicks and keystrokes. Creating robust heuristics to infer high level analytic activity from these low level events is difficult and error prone. A resulting design objective for nCompass, therefore, was to incorporate a new framework for capturing higher-level, more mean-

ingful indicators of analytic activity to be made available to user modeling services.

9 Conclusion and Future Work

Experiments have demonstrated the great potential in rich logging of analyst activity to support context sharing among users and across OSINT analysis tools to produce value-added information services. This is the key to enabling tacit collaboration, providing the analyst with improved information awareness without imposing any additional procedural or cognitive strain, and is particularly applicable to the Web 2.0 analytic tools emerging for Open Source Intelligence.

nCompass is designed to support the capture of meaningful indicators of analytic activity, and to allow a dramatically larger number of individual computational analytic capabilities, applications and services to be integrated together quickly and effectively. The nCompass SOA framework proved to be a key element in the success of researchers working to design solutions that increase analytic productivity and information value delivered to users through the use of larger frameworks of diverse, context-aware computational systems. It is by combining the analyst with shared context across multiple advanced capabilities in a system of systems that a significant improvement in analytic performance can be achieved.

Our planned next steps are to integrate with an existing analytic toolset to investigate the potential in modeling analytic workflow. At present, we have drafted an extension to the Analysis Log Event specification to support logging of workflow events in the analytic process. A prototype Analysis Log Service has been implemented to support this extended ALE specification.

An additional future direction of interest is implementing the Analysis Log Event schema in a machine-processable format, such as RDF (Resource Description Framework) or OWL (Web Ontology Language). This would enable us to explore the potential for machine reasoning over logged indicators of analytic activity.

Acknowledgments

This work was supported and monitored by the Intelligence Advanced Research Projects Activity (IARPA) and the Air Force Research Laboratory (AFRL) under contract number FA8750-06-C-0211.

The views, opinions and findings contained in this report are those of the authors and should not be construed as an official Department of Defense position, policy or decision, unless so designated by other official documentation.

The authors wish to thank the IARPA Collaborative Analyst and System Effectiveness (CASE) Program staff and AFRL staff for their support and encouragement.

References

1. ODNI Information Sharing Strategy, February 21, 2008, p. 10.
2. Proulx, P., L. Chien, R. Harper, D. Schroh, T. Kapler, D. Jonker, W. Wright, "nSpace and GeoTime - VAST 2006 Case Study", IEEE Computer Graphics and Applications, 2007.
3. Whittenton, J., Editor, "Defense Intelligence Integrators Guide (DIIG)", version 1.2, Febru-ary, 2007.
4. Erl, Thomas, "Service-Oriented Architecture - Concepts, Technology, and Design", Prentice Hall, 2005.
5. Krafzig, D., Banke, K., Slama, D., "Enterprise SOA Service-Oriented Architecture Best Practices", Prentice Hall, 2005.
6. Intelligence Advanced Research Projects Activity (IARPA) Office of Incisive Analysis, http://www.iarpa.gov/office_incisive.html/.
7. Brueckner, S., Downs, E., Hilscher, R., Yinger, A., "Self Organizing Integration of Compet-ing Reasoners for Information Matching", In Proceedings of International Workshop on Environment-Mediated Coordination in Self-Organizing and Self-Adaptive Systems (ECOSOA'08), Venice, Italy, October 20, 2008.
8. Hampson, E., Cowley, P., "Instrumenting the Intelligence Analysis Process", First Interna-tional Conference on Intelligence Analysis Methods and Tools, MITRE, McLean, VA, May 2-6, 2005.
9. General Dynamics Advanced Information Systems tag|Connect, http://www.gd-ais.com/index.cfm?acronym= tagconnect
10. Bringsjord, S., Arkoudas, K., Clark, M., Shilliday, A., Taylor, J., Schimanski, B. & Yang, Y., "Reporting on Some Logic-Based Machine Reading Research". in Etzioni, O., ed., Ma-chine Reading: Papers from the AAAI Spring Symposium (Technical Report SS-07-06), pp. 23-28, and also www.cogsci.rpi.edu/research/rair/slate.
11. Hilscher, R., Sven Brueckner, Theodore C. Belding, H. Van Dyke Parunak, "Self-Organizing Information Matching in InformANTS", SASO '07: Proceedings of the First In-ternational Conference on Self-Adaptive and Self-Organizing Systems, July, 2007.
12. Aarseth, P., M. Deligonul, J. Lehmann, L. Nezda and A. Hickl., "TASER: A Temporal and Spatial Expression Recognition and Normalization System", Proceedings of the 2005 Auto-matic Content Extraction Conference (ACE 2005), Gaithersburg, MD.
13. Freitag, D., M. Blume, J. Byrnes, R. Calmbach and R. Rohwer, "A Workbench for Rapid Development of Extraction Capabilities", In Proc. International Conference on Intelligence Analysis, 2005.
14. Moldovan, D., S. Harabagiu, R. Girju, P. Morarescu, F. Lacatusu, A. Novischi, A. Badulescu and O. Bolohan, "LCC Tools for Question Answering", Proceedings of the TREC-2002 Conference, NIST, Gaithersburg, 2002.
15. WS-I Basic Profile, http://www.ws-i.org/Profiles/BasicProfile-1_2(WGAD).html/.
16. CASE Analysis Log Service Specification, version 3.0, June 27, 2008.
17. SOAP Message Transmission Optimization Mechanism, http://www.w3.org/TR/soap12-mtom/.
18. Google Search Applicance, http://www.google.com/ enterprise/ gsa/.
19. IBM OmniFind Yahoo Edition, http://omnifind.ibm. yahoo.net/.
20. OpenSocial, http://code.google.com/apis/opensocial/.

21. Heuer, Richards J., Jr, "Chapter 8: Analysis of Competing Hypotheses", Psychology of Intelligence Analysis, Center for the Study of Intelligence, Central Intelligence Agency, 1999.
22. Morse, E., Sheppard, C., Grantham, J., Cheikes, B. and J. Boiney, "APEX System Evalua-tion", NIST and MITRE, October 30, 2008.
23. Alonso, R., and H. Li, "Model-Guided Information Discovery for Intelligence Analysis", in Proceedings of CIKM '05, Bremen, Germany, 2005.
24. Morse, E., "UMS Multi-Component Evaluation Plan", NIST, June 27, 2008.
25. Chien, L., A. Tat, P. Proulx, A. Khamisa and W. Wright, "Grand Challenge Award 2008: Support for Diverse Analytic Techniques", IEEE Visual Analytics Science and Technology VAST 2008.
26. Morse, E., "Recommender2 Evaluation: New Vectors/SET, Oculus, BAE, Fair Isaac", NIST, December 3, 2008.
27. Dey, Anind K., "Understanding and Using Context", Personal and Ubiquitous Computing, Volume 5, Issue 1, February 2001, pp. 4 - 7.
28. Chen, H. Finn, T. and A. Joshi, "An ontology for context-aware pervasive computing environments". The Knowledge Engineering Review (2003), 18, Cambridge University Press, pp. 197-207.
29. Dey, Anind K. and Gregory D. Abowd, "The Context Toolkit: Aiding the Development of Context-Aware Applications", Proc. Conf. Human Factors in Computing Systems (CHI), ACM Press, New York, 1999, pp. 434-441.
30. Hardisty, F., "GeoJabber: Finding Significant Analytic Events in Collaborative Visual Analysis Sessions", GIScience 2008, September 23-26, 2008, Park City, Utah.
31. Schmidt, A., "Implicit Human Computer Interaction Through Context", Personal Technologies 4(2-3), June 2000, pp. 191-199.

SOA Security Aspects in Web-based Architectural Design

Asadullah Shaikh, Sheeraz Ali, Nasrullah Memon, and Panagiotis Karampelas

Abstract Distributed web-based applications have been progressively increasing in number and scale over the past decades. There is an intensification of the need for security frameworks in the era of web-based applications when we refer to distributed telemedicine interoperability architectures. In contrast, Service Oriented Architecture (SOA) is gaining popularity day by day when we specially consider the web applications. SOA is playing a major role to maintain the security standards of distributed applications. This paper proposes a secure web-based architectural design by using the standards of SOA for distributed web application that maintains the interoperability and data integration through certain secure channels. We have created CRUD (Create, Read, Update, Delete) operations that has an implication on our own created web services and we propose a secure architecture that is implemented on CRUD operations.
The paper provides an extensive description of the prevention of replay attacks and a detailed explanation for applying security measures.

Asadullah Shaikh
Universitat Oberta de Catalunya, Barcelona, Spain
e-mail: ashaikh@uoc.edu

Sheeraz Ali
Cursor Software Solutions, UAE
e-mail: sheeraz@cursorsoft.net

Nasrullah Memon
Maersk Mc-Kinney Moller Institute,
University of Southern Denmark, Odense, Denmark
e-mail: memon@mmmi.sdu.dk

Panagiotis Karampelas
Hellenic American University, Athens, Greece
e-mail: pkarampelas@hau.gr

1 Introduction

Service Oriented Architecture (SOA) is gaining popularity day by day due to the fact that it is useful for making interoperable web-based applications [5, 6]. The non-secure SOA based applications create many problems for different web based applications especially concerning the security aspects. However, security in enterprise applications has not been addressed adequately. In the present situation, the security aspects in architectural designs are not considered until a serious problem occurs during the developmental stages that violates the policy. Architectural designs are always being considered as a preliminary stage of a development process, therefore, if the designing of an architecture for a system is planned to be secured, then it should not be a major problem to maintain the security during the implementation process.

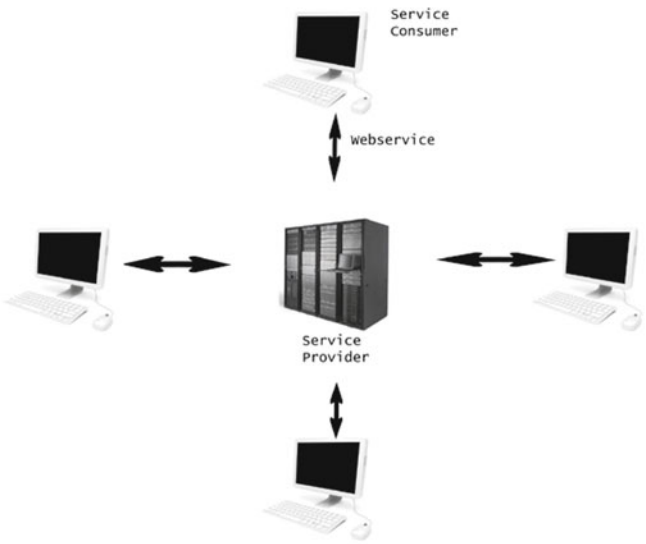


Fig. 1: Web Service Interaction

Apart from that, the current telemedicine web-based applications are being widely used in the e-health care system [14, 9], and out of them the majority of applications are distributed due to the several placement locations. These applications store the patient's history, personal details etc, which are quite confidential data. Nevertheless, there is less attention paid to the security of these heterogenous telemedicine web based applications [9]. Figure 1 refers to a normal structure of a web application interaction based on web services between the service provider and service consumer applications.

Considering all of above aspects, this paper proposes the security aspects for our developed architectural design [13] to address the security issues in order to provide a suitable solution. To configure these security aspects appropriately, we took into account several standards of security implementation over the web and multiple distributed applications. As a result, we decided to design an architecture that integrates the necessary security measures in the developmental process and is very straightforward for the developers. Our proposed secure architectural design is based on Public Key Infrastructure (PKI) security for authentication and authorization [8]. Furthermore, the proposed security aspects have an implication on CRUD operations which are web-based services that were previously implemented [13].

The rest of the paper is structured as follows. Section 1.1 introduces the concept of CRUD operations, section 1.2 defines the security using SOA and section 1.3 outlines the security problems. Furthermore, Section 2 provides an overview of web security in CRUD operations. Section 3 presents the proposed architecture. Later on, section 4 explains the experiments and results. Finally, section 5 is about previous work related to SOA security and section 6 draws on some conclusions and future work.

1.1 CRUD Operations

Our previously designed telemedicine architecture is based on SOA. We have implemented CRUD operations as a basic application function in order to interact with the database. These CRUD operations perform database operations such as data retrieval, data creation, data deletion and data updation at an application level. Table 1 briefly describes the mapping of CRUD functions with database operations.

Table 1: List of CRUD Services with Security [14]

Service Name	CRUD Type	Description	Security Type
CRUD_InsertPatientRecord	Create	Take email/MMS data from email server and then parse them and insert it into database.	WS-Security (Username Token)
CRUD_GetPatientRecord	Read	Retrieves the patient's records against the patient's personal number.	WS-Security (Username Token)
CRUD_DeletePatientRecord	Delete	Deletes the patient's records.	WS-Security (Username Token)
CRUD_UpdatePatientRecord	Update	Updates patient's record.	WS-Security (Username Token)

1.2 Security Using SOA

The SOA approach is widely used to develop the several components of web services. These services contain their own security techniques [16]. WS-Security provides a communication protocol to apply the security of web services [3]. It describes Simple Object Access Protocol (SOAP) messaging to enable security services specially in terms of integrity, message confidentiality and message authentication, and furthermore, it helps to provide encryption techniques. This kind of security provides a flexible design for security models such as Secure Sockets Layer (SSL) and Kerberos. Nevertheless, it provides security tokens, trust domains, signatures and encryption technologies. In order to exchange the secured messages using WS-Security, there will be common tokens which should be shared between requester and provider. Figure 2 describes the general structure of WS-Security.

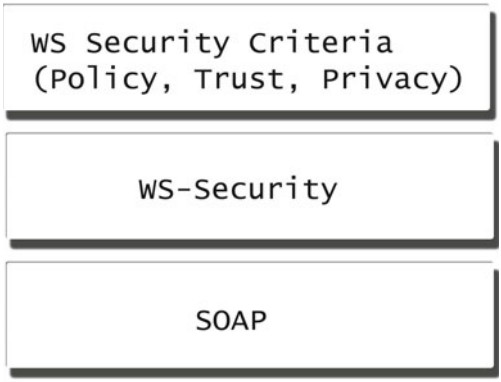


Fig. 2: WS-Security Structure [16]

1.3 Security Problems

WS-Security is quite flexible and capable, however, its configuration in real time examples is difficult for users. So far and to the best of our knowledge, the security for CRUD operations used in web services has not been discussed sufficiently. Therefore the security aspects of the following points are not considered:

- CRUD operations are developed in web services, and so what should the security measurements be in order to get the patient data?

- CRUD operations are persistent storage functions that are implemented in the form of web services [14], If the patient data is updated, how can its authentication be ensured?
- In the case of deleting a record, how can somebody be authorized to do this?

This paper presents security aspects to handle security in accordance with the CRUD operations that are implemented and accessible as a web service.

2 WS-Security in CRUD operations

CRUD is the combination of four basic operations: Create, Read, Update, and Delete which are used for permanent storage [14], and are the major components of every computer software application. Since data is the most important and valuable in any web-based application, therefore it should be transmitted securely. In our proposed work, we have provided the security for our CRUD operations using WS-Security.

3 Proposed Architecture

In this section, we describe the proposed architecture of our web services along with the CRUD security implementation. The major structure of the proposed process of SOA security for CRUD operations is divided in two external interfaces; one interface is between the nurse and the application and the other interface is between the doctor and the application. The flow of both interfaces is illustrated in Figure 3.

3.1 Interface between Nurse and Application

Initially, the first operation that needs to be performed is that the nurse sends the patient record to a telemedicine application from a cell phone with a premium phone number registered in the application. Premium numbers are the special type of cell phone numbers that are designed especially for telemedicine applications in order to process secure data transmission from nurse end to doctor end. Secondly, an application sends a random unique code for verification to the nurse end. Thirdly, the nurse replies for verification and finally, if the verification is performed successfully, then the data will be processed (update or create) to CRUD operations.

3.2 Interface between Doctor and Application

In order to maintain the security for our telemedicine application, the following steps are undertaken in the case of a doctor's interaction with application.

1. The doctor selects a CRUD operation from the given User Interface (UI) which will consume a web service. Afterwards, the running telemedicine application encrypts the message with a private key and then the secure SOAP message with encryption will be sent to the telemedicine application.
2. The security handler of a telemedicine application decrypts the message with the doctor's public key.
3. The security handler checks certain permission given to a particular doctor in order to select the CRUD operation to be performed.
4. CRUD operation is performed.

3.3 Security Measures for an Intruder

In the worst case scenario, the intruder can also try to attack our telemedicine application, therefore, if the intruder sends the operation message, the SOAP message without encryption will be sent to a telemedicine application to access the CRUD operations. Afterwards, the security handler of a telemedicine application may decrypt the message with the doctor's public key but the decryption will fail and the message will be discarded.

3.4 Security Implementation at Nurse's End

Authentication and Authorization (AA) at the nurse's end is performed in two levels. In the first level of security, the messages sent by nurse can only be accepted from premium cell phone numbers that are registered in an application. In the second level of security, once the messages are received, the telemedicine application sends a unique code to the nurse for verification, and the nurse replies to the message with the same code. The idea to introduce the second level security is to ensure that the message was sent from a premium number through a cell phone as the message can also be sent from Web2SMS or Web2MMS with any number. Therefore, a nurse needs to send the reply for verification in order to pass the secure data into CRUD operations.

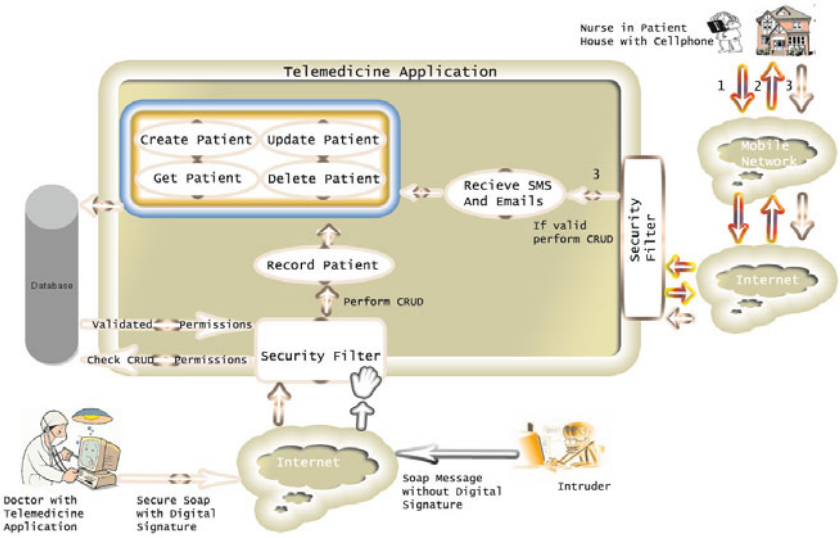


Fig. 3: Proposed Security Architectural Design

3.5 Security Implementation at the Doctor's End

We have developed the security implementation of telemedicine system architecture using Apache Axis [1], which is an open source XML based Web Service framework. We have used Apache Web Services Security for Java (WSS4J) [2] which is the implementation of the OASIS Web Services Security (WS-Security) from the Organization for the Advancement of Structured Information Standards (OASIS) Web Services Security Technical Committee (TC) [10]. WSS4J is used to sign and verify SOAP messages with WS-Security information. Furthermore, WSS4J is used for securing our CRUD web services along with the support of the Apache Axis web service framework. WSS4J generates and processes the SOAP bindings for XML Security with *XML Signature* and *XML Encryption*. It also provides the *Tokens* for username, timestamps and Security Assertion Markup Language (SAML) tokens. The security of CRUD operations is deployed with username tokens.
The configuration of the security deployment and usage is described by the implementation given in listings 1 to 5 .

In listing 1, the WSS4J handlers are added to the service deployment descriptor in the Web Service Deployment Descriptor (WSDD) file for adding the WS-Security layer to our telemedicine CRUD services. Afterwards, adding handlers, the server side deployment descriptor also defines the request and response flows. In the *RequestFlow* (Listing 1, Line 5), with every incoming re-

quest for a CRUD operation, there are two security handlers that authenticate and authorize the request. The TeleWoundServiceSecurityHandler (Listing 1, Line 21) decrypts the SOAP message with a public key of the Doctor using PKI security. Once the message is decrypted, WSDoAllReceiver (Listing 1, Line 13) verifies the username and password for authorization. Meanwhile, in the *ResponseFlow* (Listing 1, Line 18), every response is encrypted with the doctor's public key and the message is digitally signed for authentication with the telemedicine's private key.

In listing 2, a password callback class called PWCallback (Listing 2, Line 1) is created by implementing the CallbackHandler interface. This CallbackHandler is called before every CRUD operation request to check the authorization of the provided username and password. If the username and password exists in the application, then there will be the verification of permission on the selected CRUD operation (Listing 2, Line 8). For example, a user may have the access only on the CRUD operation READ, while the CRUD operation UPDATE is not permitted, then PWCallback class will not allow any action on UPDATE operation due the assigned permissions.

Listing 1: Code for Request and Response Flows

```
1  <deployment xmlns=http://xml.apache.org/axis/wsdd/ xmlns:java=
2  "http://xml.apache.org/axis/wsdd/providers/java">
3   <service name="TeleWound-wss-01" provider="java:RPC" style=
4   "document" use="literal">
5    <requestFlow>
6     <handler type="java:org.apache.axis.handlers.JAXRPCHandler">
7      <parameter name="scope" value="session"/>
8      <parameter name="className" value="TeleWoundServiceSecurityHandler"/>
9      <parameter name="keyStoreFile" value="c:\\TeleWound\\key\\server.ks"/>
10     <parameter name="trustStoreFile" value="c:\\TeleWound\\key\\server.ts"/>
11     <parameter name="certEntryAlias" value="clientkey"/>
12    </handler>
13    <handler type="java:org.apache.ws.axis.security.WSDoAllReceiver">
14     <parameter name="passwordCallbackClass" value="PWCallback"/>
15     <parameter name="action" value="UsernameToken"/>
16    </handler>
17   </requestFlow>
18   <responseFlow>
19    <handler type="java:org.apache.axis.handlers.JAXRPCHandler">
20     <parameter name="scope" value="session"/>
21     <parameter name="className" value="TeleWoundServiceSecurityHandler"/>
22     <parameter name="keyStoreFile" value="c:\\TeleWound\\key\\server.ks"/>
23     <parameter name="trustStoreFile" value="c:\\TeleWound\\key\\server.ts"/>
24     <parameter name="certEntryAlias" value="clientkey"/>
25    </handler>
26   </responseFlow>
27   <parameter name="scope" value="application"/>
28   <parameter name="className" value="TeleWound"/>
29   <parameter name="allowedMethods" value="CRUD_InsertPatientRecord"/>
30   <parameter name="allowedMethods" value="CRUD_UpdatePatientRecord"/>
31   <parameter name="allowedMethods" value="CRUD_DeletePatientRecord"/>
32   <parameter name="allowedMethods" value="CRUD_GetPatientRecord"/>
33  </service>
34 </deployment>
```

Listing 2: Code for Password Callback

```
1  public class PWCallback{
2   public void handle(Callback[] callbacks) throws
3    IOException, UnsupportedCallbackException {
4    for (int i = 0; i < callbacks.length; i++) {
5     if (callbacks[i] instanceof WSPasswordCallback) {
6      WSPasswordCallback pc = (WSPasswordCallback)callbacks[i];
7      // set the password given a username
8      if ("TeleMedicineAdmin".equals(pc.getIdentifier(){
9        // set the password
10     }}
11    else{
12     throw new UnsupportedCallbackException(callbacks[i],"Unrecognized Callback");
13   }}}
```

In Listing 3, TeleWoundServiceSecurityHandler class is securing the message by encryption and decryption using PKI security. The message is digitally signed to authenticate the provider of the message. With every incoming CRUD request, handleRequest() method (Listing 3, Line 5) is called to decrypt the SOAP message with sender's public key. Therefore, for every outgoing response, handleResponse() method (Listing 3, Line 23) is called for encrypting the SOAP message and to digitally sign the message.

Listing 3: Code for Authentication

```
1  Public class TeleWoundServiceSecurityHandler implements Handler {
2    private String keyStoreFile, keyStoreType, keyStorePassword ,
3          keyEntryAlias, keyEntryPassword, trustStoreFile,
4          trustStoreType, trustStorePassword, certEntryAlias;
5    public boolean handleRequest(MessageContext context) {
6      try {
7        SOAPMessageContext soapContext = (SOAPMessageContext)context;
8        SOAPMessage soapMessage = soapContext.getMessage();
9        Document doc = SOAPUtility.toDocument(soapMessage);
10       Utility.decrypt(doc, keyStoreFile, keyStoreType,
11             keyStorePassword, keyEntryAlias, keyEntryPassword);
12       Utility.verify(doc, trustStoreFile, trustStoreType,
13             trustStorePassword);
14       Utility.cleanup(doc);
15       soapMessage = SOAPUtility.toSOAPMessage(doc);
16       soapContext.setMessage(soapMessage);
17     } catch (Exception e){
18        System.err.println("handleRequest Exception:" + e);
19        return false;
20     }
21     return true;
22   }
23   public boolean handleResponse(MessageContext context) {
24     try {
25       SOAPMessageContext soapContext = (SOAPMessageContext)context;
26       SOAPMessage soapMessage = soapContext.getMessage();
27       Document doc = SOAPUtility.toDocument(soapMessage);
28       Utility.sign(doc, keyStoreFile, keyStoreType,
29            keyStorePassword, keyEntryAlias, keyEntryPassword);
30       Utility.encrypt(doc, trustStoreFile, trustStoreType,
31            trustStorePassword, certEntryAlias);
32       soapMessage = SOAPUtility.toSOAPMessage(doc);
33       soapContext.setMessage(soapMessage);
34     } catch (Exception e){
35       System.err.println("handleResponse Exception:" + e);
36       return false;
37     }
38     return true;
39   }
40   public boolean handleFault(MessageContext context) {
41     return true;
42   }
43   public void init(HandlerInfo config) {
44     Map configProps = config.getHandlerConfig();
45     keyStoreFile = (String)configProps.get("keyStoreFile");
46     keyStoreType = (String) configProps.get("keyStoreType");
47     keyStorePassword = (String) configProps.get("keyStorePassword");
48     keyEntryAlias = (String) configProps.get("keyEntryAlias");
49     keyEntryPassword = (String) configProps.get("keyEntryPassword");
50     trustStoreFile = (String)configProps.get("trustStoreFile");
51     trustStoreType = (String) configProps.get("trustStoreType");
52     trustStorePassword = (String) configProps.get("trustStorePassword");
53     certEntryAlias = (String) configProps.get("certEntryAlias");
54   }}
```

Listing 4: Code for Client Request

```
<deployment xmlns="http://xml.apache.org/axis/wsdd/"
java="http://xml.apache.org/axis/wsdd/providers/java">
 <transport name="http" pivot="java:org.apache.axis.
_transport.http.HTTPSender">
  <globalconfiguration>
   <requestFlow>
    <handler type="java:org.apache.axis.handlers.JAXRPCHandler">
     <parameter name="scope" value="session"/>
     <parameter name="className" value="TeleWoundServiceSecurityHandler"/>
     <parameter name="keyStoreFile" value="c:\\TeleWound\\key\\server.ks"/>
     <parameter name="trustStoreFile" value="c:\\TeleWound\\key\\server.ts"/>
     <parameter name="certEntryAlias" value="clientkey"/>
    </handler>
  </globalconfiguration>
 </transport>
</transport>
</deployment>
```

3.6 Prevention from Replay Attack

Replay attacks are also called "man-in-the-middle attack" where a hacker makes independent connections in order to break the barrier of an application. As discussed before, we have implemented the security on a bottom level which is efficient, less expensive and easy to maintain for developers and designers[15]. One of the major factors in our proposed architecture is the prevention from replay attacks where an attacker can access our encrypted SOAP messages to extract the patient's confessional data. Figure 4 illustrates the scenario in which the hacker/intruder tries to breach encrypted security messages in order to reach at the telemedicine web-based application. To prevent this, we have used PKI and digital signatures for encryption, authentication and authorization. However, the authentication is not limited to only that but also the digitally signed message can be recorded through several capturing techniques which can be resend. We call this a Replay Attack.

Replay attacks can be prevented by making each SOAP message unique. There are different type of techniques used for uniqueness of SOAP message like a number or bit string used only once (nonce), time stamping etc. Fortunately, WSS4J provides time stamping for each SOAP message to prevent replay attack easily. In order to avoid these types of attacks, we have used timestamping in our security architecture. Timestamping is a chain of characters, pointing the data or time at which an event is occurred. In timestamping, each event is recorded by a computer. Listing 5 represents the code used in our telemedicine application to avoid the replay attacks through timestamping. We have set the maximum duration of 10 seconds of each timestamp. Figure 5 shows an over all view after prevention from replay attacks.

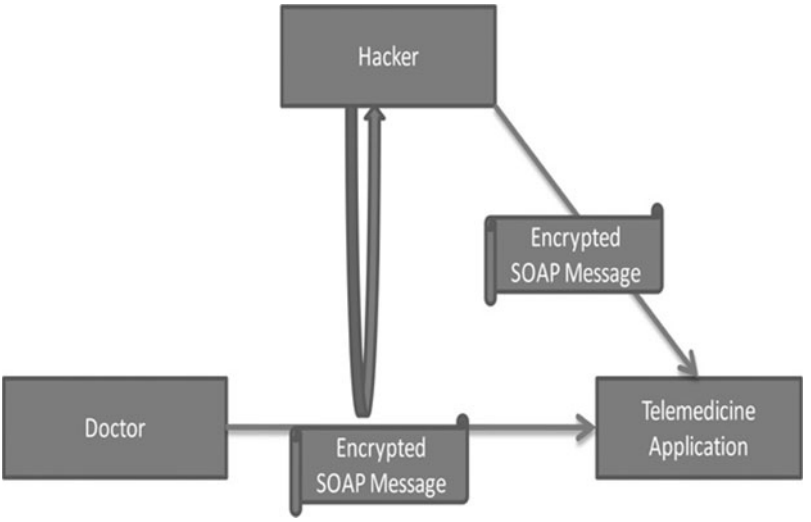


Fig. 4: Proposed Security Architectural Design

Listing 5: Code for Timestamping

```
<S:Envelope    xmlns:S="http://www.w3.org/2001/12/envelope"
   xmlns:wsu="http://schemas.xmlsoap.org/ws/2002/07/utility">
<S:Header>
                    <wsu:Timestamp>
                          <wsu:Created>2009-09-15T06:27:00Z</wsu:Created>
                          <wsu:Expires>2009-09-15T06:37:00Z</wsu:Expires>
                          <wsu:Received Actor="http://telemedicine.com/" Delay="60000">
                    2009-09-15T06:30:00Z
                          </wsu:Received>
                    </wsu:Timestamp>
                    <!--Other Header Details-->
          </S:Header>
          <S:Body>
          <!--SOAP Message here-->
          </S:Body>
</S:Envelope>
```

4 Experiments and Results

In this section, we present the context in which the experiment examines the assumptions that are required to maintain the consistency and dynamics of the proposed security architecture. Table 2 summarizes the experimental results obtained by the implementation of our proposed technique. Column *Actor*, represents an actor who is the user of the telemedicine application, column *CRUD Operations(Encode)* describes the operation called by an actor who has certain rights. Column *CRUD Data* displays patient's data, column *Encryption + Signature* describes the encrypted CRUD data along with the signature and the column *Security Type* shows the type of security implemented on prescribed actor depending on permissions. Finally, the column

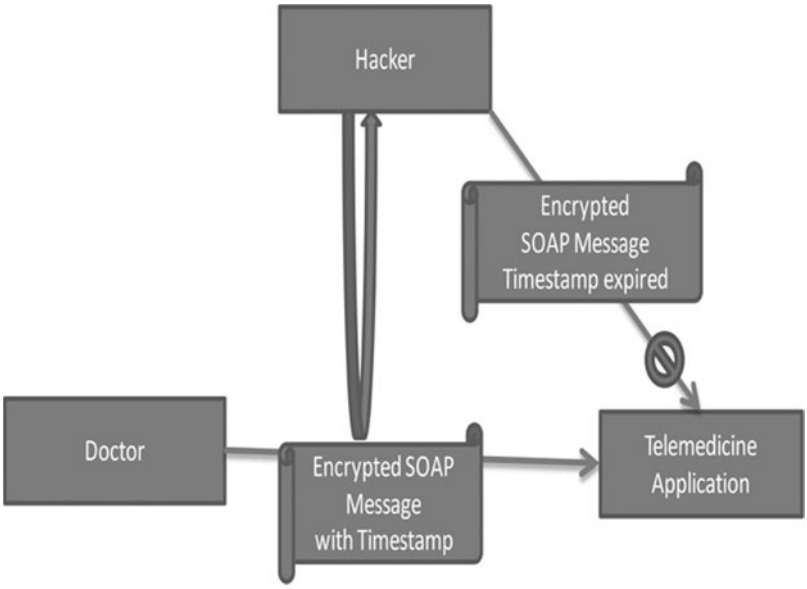


Fig. 5: Proposed Security Architectural Design

Table 2: Results of proposed security Architecture

Actor	CRUD Operations (Encode)	CRUD Data	Encryption + Signature	Security Type	Decryption	Receiving Message (decode)
Nurse	Create,Update Patient Record	Patient Data	N/A	Premium Number +Verification Code	N/A	1. Patient Record 2. Verification Code
Doctor	Update,Delete,Read Patient Record	Patient Data	BN89bOi + 978654	PKI Encryption+ Digital Signature	Patient Data	CRUD Operation Perfomed
Intruder	Create,Update,Delete Read,Patient Record	Patient Data	Any Encryption or Signature	Invalid PKI Encryption +Digital Signature	No Data Received due to wrong Encryption + Signature	CRUD Operation Failed

Decryption shows the patient's decrypted data if and only if all the security steps become successful and the column *Receiving Message(Decode)* shows a message once the decryption has failed or passed. All these results provide an overview of our developed security over CRUD operations.

4.1 Security Scenarios

For the sake of brevity and without loss of generality, in this section we consider different scenarios of security. These scenarios are based on obtained result from our prototype (Table 3):

Table 3: Scenarios for Timestamping

	SOAP Message	Encrypted SOAP Message	Message Created at	Message Expired at	Current Time	End Result
Without Timstamping (Doctor)	getPatient("abc")	Encrypted with digital signature	—	—	—	Message executed successfully
Without Timstamping (Intruder)	Encrypted Message recorded to replay	(Recorded) Encrypted with digital signature	—	—	—	Recorded Message executed successfully
With Timstamping (Doctor)	getPatient("abc") with timestamp	Encrypted with digital signature	2009-09-15T06:27:00Z	2009-09-15T06:37:00Z	2009-09-15T06:32:00Z	Message executed successfully before expiry time
With Timstamping (Intruder)	Encrypted Message recorded to replay	(Recorded) Encrypted with digital signature	2009-09-15T06:27:00Z	2009-09-15T06:37:00Z	2009-09-15T06:48:00Z	Recorded Message rejected as time is expired.

Scenario 1 (No Security Implemention): In simple SOA implementation without the security, there is a risk that intruder can access the confidential information by simple capture of network traffic.

Scenario 2 (Message is Encrypted to Protect Confidential Data of a Patient): To avoid the access of confidential information to intruder, we have implemented the encryption to secure the SOAP messages during communication between Telemedicine System and its client application for doctors. However, intruder can use the same encryption mechanism to generate SOAP messages and exploit the functionality of Telemedicine System.

Scenario 3 (Message is Signed with the Digital Signature for Authentication and Authorization): For authentication and authorization of the doctor, we have used PKI infrastructure to digitally sign the SOAP messages to ensure that the messages are coming from the real and registered doctor, who has the access to the system. Although intruder cannot decrypt and change the message but still there is a chance to capture and record the signed encrypted message and replay the same message to exploit the functionality of the system. We call this replay attack.

Scenario 4 (The Message is Time stamped to Prevent Replay Attack): To prevent the system from replay attack, we have used the time stamping. In time stamping, we append the creation and expiry time with each message. As recording and replaying of message take some time therefore, within that time frame, the message will be expired, thus the expired messages will be rejected by the Telemedicine System.

5 Related Work

In this section, we discuss the existing work on SOA Security. Most of the work in this area is done for SOA security specification. Their goal of implementing the security is to achieve the authentication, authorization and so on for web services. However, research work on the CRUD operations using WS-Security is hardly found in the literature.

Phan, Cecilia [11] addressed the security challenges for SOA. The author described the problems raised from XML which is not secure enough and causes problems in security protocol. They also presented certain strategies to cope with vulnerabilities against attacks and other security policy consideration.

Larrucea [7] proposed an approach describing a holistic view of a SOA environment. In this research, ISOAS framework allows functionality criterions of security policies with service specification that allows the definition of functional and non functional components in coherent way and is dependent on the metamodel. This effort is implemented in Eclipse, and it is due to that, that it is an open approach. Apart from that, their approach is aligned with OMG standards.

Satoh F et al. [12] discussed a process of security configuration that defines the responsibilities of developers. In this end-to-end SOA security configuration, several kinds of information are needed such as requirements, platform information and so on. Due to that, they defined the roles of developers during the development phase. SOA security is complex therefore the domain federation is considered in this research. In general, they contribute to the correct configuration to reduce the workload of developers.

Robert Bunge et al. [4] proposed an operational framework of a network administrator using SOA network security. In this research, they characterize the steps in SOA network security in order to collect the information regarding threats and SOA deployments. Furthermore, they collect the SOA security efforts. As a result, by considering the factors of SOA network security, they provide recommendations for dealing with the XML network traffic for SOA applications. The proposed approach is filtered to inspect XML at the network's level. Their framework contributes to secure SOA design by clarifying the duties of network administrators and software engineers using XML-based services.

Yamany H, Capretz M [16] described an intelligent security service that is embedded in a framework to secure web services in SOA. This framework is designed to interact with authentication to run the authentication process and it also helps to secure a possible web attack. An SOA environment holds several security environments that interact through multiple channels. In their work, they have examined the security service layer and message security layer.

All the above work presented so far is not similar to ours, because of implementing security over CRUD operations. If the CRUD operations are secured enough, then there is no need to apply high level security which is a definitely a complex task. Our CRUD operations are interacting with created web services, therefore if we apply the security on CRUD, the web-service will also be secured simultaneously. However, we have designed and implemented a system architecture that represents the scenario of security standards by considering CRUD operations along with web services.

6 Conclusion and Future Work

In this work, we have proposed an architectural design by considering the security aspects for our designed CRUD operations using SOA. We believe that SOA has multiple solutions of web services. The core use of CRUD operations is to fetch, update, delete, read the data from a perspective database, therefore if CRUD is secure enough, then there is no such need to implement the high level security. In our designed architecture, communication is done through SOAP messages and we have implemented WSS4J and PKI security in order to protect SOAP headers. It creates the efficiency of the security process and prevents web attacks. As a future work, the application of similar efficient security techniques can be explored in cloud computing.

References

1. Apache. Apache Axis, Accessed December.24,2009 [Online]. http://ws.apache.org/axis.
2. Apache. Apache WSS4J, Accessed December.24,2009 [Online]. http://ws.apache.org/wss4j.
3. G. S. Bob Atkinson, et al. Web services security (ws-security), copyright 2002-2002 international business machines corporation, microsoft corporation, Accessed December.24,2009 [Online]. http://www.cgisecurity.com/ws/ws-secure.pdf.
4. R. Bunge, S. Chung, B. Endicott-Popovsky, and D. McLane. An operational framework for service oriented architecture network security. In *HICSS '08: Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, page 312, Washington, DC, USA, 2008. IEEE Computer Society.
5. X. Chen, G. Huang, and H. Mei. Towards automatic verification of web-based soa applications. In *APWeb*, pages 528–536, 2008.
6. N. A. Delessy and E. B. Fernandez. A pattern-driven security process for soa applications. *Availability, Reliability and Security, International Conference on*, 0:416–421, 2008.
7. X. Larrucea and R. Alonso. ISOAS: Through an independent SOA security specification. In *Proceedings of the Seventh International Conference on Composition-Based Software Systems (ICCBSS 2008)*, pages 92–100. IEEE Computer Society, 2008.
8. MSDN. X.509 technical supplement, Accessed December.24,2009 [Online]. http://msdn.microsoft.com/en-us/library/aa480610.aspx.
9. W. M. Omar and A. Taleb-Bendiab. Service oriented architecture for e-health support services based on grid computing over. *Services Computing, IEEE International Conference on*, 0:135–142, 2006.
10. Organization for the Advancement of Structured Information Standards(OASIS). Web services security technical committee, Accessed April.3,2009 [Online]. http://www.oasis-open.org/committees/tchome.php?wgabbrev=wss.
11. C. Phan. Service Oriented Architecture (SOA)-Security Challenges and Mitigation Strategies. In *IEEE Military Communications Conference, 2007. MILCOM 2007*, pages 1–7, 2007.
12. F. Satoh, Y. Nakamura, N. K. Mukhi, M. Tatsubori, and K. Ono. Methodology and tools for end-to-end soa security configurations. In *SERVICES '08: Proceedings of the 2008 IEEE Congress on Services - Part I*, pages 307–314, Washington, DC, USA, 2008. IEEE Computer Society.

13. A. Shaikh, M. Memon, N. Memon, and M. Misbahuddin. The role of service oriented architecture in telemedicine healthcare system. *Complex, Intelligent and Software Intensive Systems, International Conference*, 0:208–214, 2009.
14. A. Shaikh, M. Misbahuddin, and M. S. Memon. A system design for a telemedicine health care system. In *IMTIC*, pages 295–305, 2008.
15. A. Shaikh, A. Soomro, S. Ali, and N. Memon. The security aspects in web-based architectural design using service oriented architecture. In *13th International Conference on Information Visualisation, IV 09, 15-17 July 2009, Barcelona, Spain*, pages 461–466, 2009.
16. H. Yamany and M. Capretz. Use of Data Mining to Enhance Security for SOA. In *Convergence and Hybrid Information Technology, 2008. ICCIT'08. Third International Conference on*, volume 1, 2008.