INTERNATIONAL SERIES IN OPERATIONS RESEARCH AND MANAGEMENT SCIENCE
ADVANCING THE STATE-OF-THE-ART

Volume Contributors
Hillel Bar-Gera
David Boyce
S.T.S. Bukkapatnam
Elaine Chang
Terry L. Friesz
Sean P. Gorman
Günter Haag
Kingsley E. Haynes
José Holguín-Veras
Soulaymane Kachani
Rajendra G. Kulkarni
S.R.T. Kumara
Shaun Lichter
Zugang Liu
Reetabrata Mookherjee
Anna Nagurney
Peter Nijkamp
Roberto Patuelli
Srinivas Peeta
Georgia Perakis
Aura Reggiani
Carine Simon
Roger R. Stough
Satish V. Ukkusuri
Sandra Vinciguerra
Athanasios Ziliaskopoulos

Network Science, Nonlinear Science and Infrastructure Systems

Edited by
Terry L. Friesz

Springer

NETWORK SCIENCE, NONLINEAR SCIENCE AND INFRASTRUCTURE SYSTEMS

Recent titles in the **INTERNATIONAL SERIES IN OPERATIONS RESEARCH & MANAGEMENT SCIENCE**

Frederick S. Hillier, Series Editor, *Stanford University*

Gass & Assad/ *AN ANNOTATED TIMELINE OF OPERATIONS RESEARCH: An Informal History*

Greenberg/ *TUTORIALS ON EMERGING METHODOLOGIES AND APPLICATIONS IN OPERATIONS RESEARCH*

Weber/ *UNCERTAINTY IN THE ELECTRIC POWER INDUSTRY: Methods and Models for Decision Support*

Figueira, Greco & Ehrgott/ *MULTIPLE CRITERIA DECISION ANALYSIS: State of the Art Surveys*

Reveliotis/ *REAL-TIME MANAGEMENT OF RESOURCE ALLOCATIONS SYSTEMS: A Discrete Event Systems Approach*

Kall & Mayer/ *STOCHASTIC LINEAR PROGRAMMING: Models, Theory, and Computation*

Sethi, Yan & Zhang/ *INVENTORY AND SUPPLY CHAIN MANAGEMENT WITH FORECAST UPDATES*

Cox/ *QUANTITATIVE HEALTH RISK ANALYSIS METHODS: Modeling the Human Health Impacts of Antibiotics Used in Food Animals*

Ching & Ng/ *MARKOV CHAINS: Models, Algorithms and Applications*

Li & Sun/ *NONLINEAR INTEGER PROGRAMMING*

Kaliszewski/ *SOFT COMPUTING FOR COMPLEX MULTIPLE CRITERIA DECISION MAKING*

Bouyssou et al/ *EVALUATION AND DECISION MODELS WITH MULTIPLE CRITERIA: Stepping stones for the analyst*

Blecker & Friedrich/ *MASS CUSTOMIZATION: Challenges and Solutions*

Appa, Pitsoulis & Williams/ *HANDBOOK ON MODELLING FOR DISCRETE OPTIMIZATION*

Herrmann/ *HANDBOOK OF PRODUCTION SCHEDULING*

Axsäter/ *INVENTORY CONTROL, 2nd Ed.*

Hall/ *PATIENT FLOW: Reducing Delay in Healthcare Delivery*

Józefowska & Węglarz/ *PERSPECTIVES IN MODERN PROJECT SCHEDULING*

Tian & Zhang/ *VACATION QUEUEING MODELS: Theory and Applications*

Yan, Yin & Zhang/ *STOCHASTIC PROCESSES, OPTIMIZATION, AND CONTROL THEORY APPLICATIONS IN FINANCIAL ENGINEERING, QUEUEING NETWORKS, AND MANUFACTURING SYSTEMS*

Saaty & Vargas/ *DECISION MAKING WITH THE ANALYTIC NETWORK PROCESS: Economic, Political, Social & Technological Applications w. Benefits, Opportunities, Costs & Risks*

Yu/ *TECHNOLOGY PORTFOLIO PLANNING AND MANAGEMENT: Practical Concepts and Tools*

Kandiller/ *PRINCIPLES OF MATHEMATICS IN OPERATIONS RESEARCH*

Lee & Lee/ *BUILDING SUPPLY CHAIN EXCELLENCE IN EMERGING ECONOMIES*

Weintraub/ *MANAGEMENT OF NATURAL RESOURCES: A Handbook of Operations Research Models, Algorithms, and Implementations*

Hooker/ *INTEGRATED METHODS FOR OPTIMIZATION*

Dawande et al/ *THROUGHPUT OPTIMIZATION IN ROBOTIC CELLS*

**** A list of the early publications in the series is at the end of the book ****

NETWORK SCIENCE, NONLINEAR SCIENCE AND INFRASTRUCTURE SYSTEMS

Edited by

Terry L. Friesz

Springer

Terry L. Friesz
Pennsylvania State University
University Park, PA, USA

Library of Congress Control Number: 2007921680

ISBN-13: 978-0-387-71080-8 (HB)   ISBN-13: 978-0-387-71134-8 (e-book)

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Contents

Preface vii

Chapter 1: A Revolution in Infrastructure Network Research and Engineering?
Terry L. Friesz 1

Chapter 2: Networks and Dynamics: The Structure of the World We Live In
Shaun Lichter and Terry L. Friesz 7

Chapter 3: Modeling Large Scale and Complex Infrastructure Systems as Computable Games
Terry L. Friesz, Reetabrata Mookherjee, and Srinivas Peeta 53

Chapter 4: Dynamic Competition on Networks: Differential Variational Inequalities, Limited Warfare and Internet Vulnerability
Terry L. Friesz 77

Chapter 5: Characterization and Monitoring of Nonlinear Dynamics and Chaos in Manufacturing Enterprise Systems
S.R.T. Kumara and S.T.S. Bukkapatnam 99

Chapter 6: Evolutionary Traffic Flow Landscapes: A Fitness Approach for ITS Management
Kingsley E. Haynes, Rajendra G. Kulkarni, and Roger R. Stough 123

Chapter 7: Network Connectivity Models: An Overview and Empirical Applications
Aura Reggiani and Sandra Vinciguerra 147

Chapter 8: An Application of Complex Network Theory to German Commuting Patterns
Sean P. Gorman, Roberto Patuelli, Aura Reggiani, Peter Nijkamp, Rajendra Kulkarni, and Günter Haag 167

Chapter 9: Assessing Critical Components in Transportation Systems: Economic Models and Complex Network Science Approaches
Satish V. Ukkusuri and José Holguín-Veras 187

Chapter 10: A Simulation-Based Dynamic Intermodal Network Equilibrium Algorithm
Elaine Chang and Athanasios Ziliaskopoulos 201

Chapter 11: Modeling the Transient Nature of Dynamic Pricing with Demand Learning in a Competitive Environment
Soulaymane Kachani, Georgia Perakis, and Carine Simon 223

Chapter 12: An Evolutionary Variational Inequality Formulation of Supply Chain Networks with Time-Varying Demands
Anna Nagurney and Zugang Liu 269

Chapter 13: Some Amazing Properties of Road Traffic Network Equilibria
Hillel Bar-Gera and David Boyce 305

Index 337

Preface

This book presents a record of a U.S. National Science Foundation Workshop held at the Pennsylvania State University in 2005. More detail about the workshop may be found in Chapter 1.

Each chapter is largely self-contained. The special strength of the book is the fact that it offers both introductory and advanced essays on each of its main topics: network science, nonlinear science, and dynamic game theory – as well as the application of those disciplines to infrastructure systems.

Chapter 1

A Revolution in Infrastructure Network Research and Engineering?

Terry L. Friesz

Harold and Inge Marcus Chaired Professor of Industrial Engineering, The Pennsylvania State University, University Park, PA 16802, USA, E-mail: tfiesz@psu.edu

The Emergence of Network Science

In the past five years a number of physicists, Internet theorists, social scientists and specialists in dynamic systems and complexity have made major strides in the development of a general theory of networks. This theory and its empirical foundation – increasingly referred to as network science – seeks to explain why networks appear and how they grow and evolve. In fact network science now has recognition among non-scientists, thanks in part to a recent popular book by Barabási (2002) that marks the coalescence and emergence of network science as a field of scientific inquiry. The agreement of other scholars with Barabási 's assertion that a new science of networks is emerging is documented in the recent comprehensive review of complex networks by Newman (2003). The review by Newman explicitly recognizes the interdisciplinary nature of network science research and the major contributions to it made by social scientists.

In brief, network science has uncovered immense similarities in diverse networks – encompassing but not limited to social networks, the Internet, roadways and terrorist networks. One of its main insights is that sub-networks are sometimes weakly coupled to one another while the sub-networks themselves are internally fully connected, which topology frequently goes hand-in-hand with the appearance of dominant hubs that have a disproportionate number of incident links. These generalizations have been established through some very ingenious experiments and mathematical analyses that challenge in a very fundamental way the old ideas of random networks and Brownian motion.

Some of the most important contributions to network science have been made by social scientists studying human activities expressed as so-called social networks. Many of the same general properties identified for physical networks – such as small worldness and scale freeness – have been discovered within social networks.

Engineering Relevance

An unanswered question is whether the results of network science will prove useful in engineering applications. A specific version of this question is whether network science can be applied to the study of infrastructure networks. Especially interesting is the potential of fusing social network models which describe human activity patterns in great detail with traditional infrastructure network models to create a new generation of network models that provide greater potential for understanding the interaction of social and infrastructure networks.

A central issue in the fusion of social network and traditional physical infrastructure network models is that of computability, since the fusion will create networks of immense size.

The Workshop

The U.S. National Science Foundation provided funds to conduct a small workshop on the relevance of network science and allied disciplines to infrastructure system research and engineering. At the workshop, Peter Dodds and Lou Pecora presented seminars on network science and nonlinear science; additionally the attendees were given a list of network science and nonlinear science readings to be completed prior to the workshop.

The workshop was mainly attended by scholars who have been highly active in one or more aspects of infrastructure systems research. In particular, experts on dynamic games, optimization, nonlinear science, sociology, spatial economics, transportation systems, GI-science, and regional science were present. The delegates who attended are:

1. Terry L. Friesz, Penn State (convener)
2. Aura Reggiani, University of Bologna (Italy)
3. Kingsley Haynes, George Mason University
4. David Boyce, University of Illinois at Chicago
5. Peter Dodds, Columbia University
6. Jose Holguin-Veras, RPI
7. Srinivas Peeta, Purdue University
8. Georgia Perakis, MIT
9. Anna Nagurney, University of Massachusetts

10. Sean Gorman, George Mason University
11. Roberto Patuelli, George Mason University
12. Louis Pecora, Office of Naval Research
13. Elaine Chang, University of South Florida

The goal of the workshop – which took place on 9, 10 and 11 May 2005 – was to focus on how network science, and to a lesser extent nonlinear science, can assist in constructing computable models that combine detailed social networks with network flow models traditionally used to study physical infrastructure.

Even though the dominant physical infrastructure considered was transportation, the delegates who attended and the authors of papers appearing in this book possess considerable experience in modeling other forms of infrastructure – including telecommunication systems, data networks, and energy distribution networks.

The individual papers that appear as chapters of this book are of three principal types:

1. those summarizing the literature on network science, nonlinear science and game theory;
2. those offering original applications of network science, nonlinear science and game theory to infrastructure systems; and
3. those illustrating how infrastructure networks have been modeled historically, including the identification of assumptions that limit their utility as decision support tools.

All included papers are original manuscripts not previously published. It is hoped that this compendium quite literally offers something for everyone interested in innovating new paradigms for the study of infrastructure systems.

Findings of the Workshop

The workshop included a discussion of each manuscript presented, led by designated discussants. In addition, after the final paper was presented, an entire afternoon was used for a free form discussion of infrastructure modeling and the potential for network science and nonlinear science to influence future infrastructure modeling efforts.

The aforementioned free-form discussion led to the following major findings of the workshop:

1. Network science may be viewed as an effort to find universal principles and laws that apply to virtually all networks, be they social or technological.
2. Network science focuses on the phenomenon of emergence, which is largely ignored by infrastructure engineers.

3. Infrastructure engineers need to look in much greater depth at the phenomenon of emergence so that structural changes and phase shifts in engineered network systems may be anticipated and guided rather than reacted to.
4. Computational considerations are paramount to infrastructure engineering, for two reasons: (a) Most infrastructure systems are very large and complex, and thereby often frustrate qualitative mathematical analysis; and (b) The practice of infrastructure engineering includes planning and forecasting; therefore it is necessary to compute – with reasonable accuracy – the future infrastructure activity levels for relevant political/governmental decision-making environments.
5. As a discipline network science has much to learn from the computational prowess developed in infrastructure network engineering that allows equilibria and transient behaviors of extremely large networks to be calculated without resort to simulation.
6. Nonlinear science has proven to be a powerful perspective for studying dynamical systems but as of yet there are no infrastructure network models that have been thoroughly analyzed from the perspective of nonlinear science. In particular dynamic models of network infrastructure design take the form of optimal control problems for which a solution may be represented as a system of simultaneous state and adjoint differential equations that can be studied to identify complexity and emergence, especially the possibility of chaos and strange attractors. Such an analysis has not been attempted and represents a 'hole' in the infrastructure network design literature.
7. Nonlinear science as a discipline has grown up divorced from dynamic game theory although the frontier in modeling infrastructure networks is widely considered to be that of dynamic games played out by agents active on infrastructure networks. This missing connection between dynamic game theory and nonlinear science has begun to be addressed by some infrastructure engineers seeking to build computable dynamic game theoretic models of specific infrastructures (most notably dynamic traffic assignment models for automobile traffic), yet the skills and insights of applied mathematicians are very greatly needed to fill this 'hole', as well.
8. Cybertechnology – a name and notion coined by the National Science Foundation to describe the use of information technology to enhance the effectiveness of the traditional infrastructure of scholarship – and cybernetworks comprised of that technology must now be considered when modeling infrastructure systems. The findings of network science concerning social networks seem potentially relevant to the study of cybernetworks.

The Future

A final finding of the workshop and one which seems appropriate to mark the end of this initial chapter is: the momentum gained in this workshop needs to be maintained by subsequent meetings and workshops for there indeed seems to be a rising interest in the holistic view of infrastructure which is at the heart of network science. As part of that holistic view, nonlinear science should receive greater emphasis from infrastructure engineers, as there is virtual unanimity among them that, in the so-called information age, the study of infrastructure involves dynamics.

References

A.-L. Barabàsi, (2002), *Linked: The "Network" Science of Networks,* Perseus Books Group.

M.E.J. Newman, (2003), "The Structure and Function of Complex Networks", *SIAM Review* **45**, 167–256.

Chapter 2

Networks and Dynamics: The Structure of the World We Live In

Shaun Lichter[1] and Terry L. Friesz[2]

[1] *Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802, USA, E-mail: lichter@psu.edu*

[2] *Harold and Inge Marcus Chaired Professor of Industrial Engineering, The Pennsylvania State University, University Park, PA 16802, USA, E-mail: tfriesz@psu.edu*

Abstract Many complex networks of systems have structures with similar topological properties: e.g. clustering, small world effect, and scale free structure. The dynamics of a system – either static or dynamic – are effected by the topological structure of the underlying network. Some examples of static and dynamic systems that act on networks include social, epidemiological, and transportation systems. This chapter gives an introduction to the analysis of nonlinear dynamics as it applies to such systems.

Keywords: network science; complex networks; nonlinear dynamics

1. Introduction

In this chapter we seek to explain the basic knowledge that is needed to understand and apply the results of network science and nonlinear science. Familiarity with the material of this chapter will allow the reader to better understand subsequent chapters of this book. The tone of this chapter is quite informal, and the mathematical background required is that of elementary calculus, an introduction to ordinary differential equations, and a working knowledge of the essentials of graph theory. It is also desirable to have some familiarity with transportation networks and the key issues arising in the design and operation of infrastructure systems – although this last mentioned familiarity is not essential.

2. Network Science

Everyday, each one of us travels to work. Some people wake up at dawn in order to make it to their New York or Los Angeles job on time. These people have to leave at the break of day in order to beat the traffic. If there is an accident on the highway on the way to work, it will take longer to get to other side of the highway or they may even have to take an alternate route. If they have to travel over a bridge, alternate routes may be very inconvenient. Some roads may have construction or a street light may go out, all of which slow down the traffic passing by. It may be curious to note that sometimes even if nothing has gone wrong, that is, with no accidents and no construction, some roads may still be highly congested simply because too many people are trying to use the same road.

Why is it that some roads may break down and very few people even notice and yet other roads can break down and the effects may be catastrophic? It may not be at the front of the average driver's mind, but on the way to work, we each travel through a transportation network. This network is a web of streets, highways, and intersections that we have to travel through to get from our starting point (home) to our destination (work). Networks like this have various topological properties, these are characteristics of the structure of the network. Learning about the structure of the network will help us to be able to answer questions like "The failure of which roads would cause the greatest harm to the network?" Maybe more importantly, "What can we do to prevent catastrophe?" Network Science is the study of the topology of networks and its role in the functionality of the network.

2.1 Social Networks

Networks have been used quite a bit to model social relationships and this easy model will help acquaint the reader with networks. In these models, each person is represented as a vertex or a node. A relationship between two people is represented as an arc or edge between their nodes. For example, in the very small social network in Figure 1 Alice, Bob, and Charlie are all represented as vertices. Since Alice is connected to Bob and Charlie via an arc, this means that Alice knows Bob and Charlie. However, since Bob and Charlie are not connected through an arc, they do not know each other.

From this simple representation of a social network, new questions may come to mind. Who is connected to the most number of people? How many people would you have to go through in order to connect a particular pair of people? Which two people are the furthest away from one another? Which people are the most crucial to the network? To begin to answer these questions, some basic definitions will help clarify the language of the following pages.

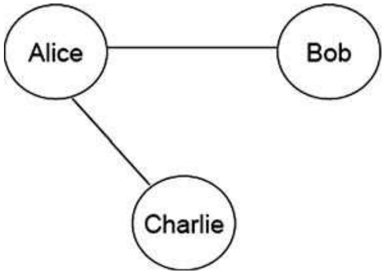


Figure 1. Social Network.

2.2 Basic Definitions

We shall have cause to employ the following notions central to graph theory and network science:

vertex	The connection of arcs, often referred to as a node.
edge	The link or arc connecting two vertices.
degree of a vertex	The number of edges connected to the vertex.
component of a vertex	The set of vertices reached from the given vertex.
geodesic path	The shortest path between two vertices in a network.
diameter	The longest geodesic path of a network.

The definitions presented above have been taken from the review article by Newman (2003).

2.3 Graph Theory is Born

The Pregel river splits into two paths around the island Kneiphof in the city of Konigsberg, Prussia (now Kaliningrad, Russia). In Konigsberg, the people of the city used to spend their Sunday afternoons walking over the seven bridges that connected the island to the pieces of land surrounding the island. The question of whether there was a path that allowed a person to walk over all seven bridges exactly once arose among the patrons of coffee shops near the bridges; it later became known in the scholarly literature as "the seven bridges of Konigsberg" problem. Figure 2, gives an abstract depiction of the configuration of the Konigsberg bridges at issue. The historical record also indicates this question soon came to the attention of Leonhard Euler (1707–1783), who isolated the underlying graph to obtain a depiction like that of Figure 3 taken from Barabási (2002).

How did Euler approach this problem? Euler reasoned that if a traveler wants to pass through a node (land mass), then he has to enter and leave, so there must be an even number of arcs for each node that the traveler passes through. Similarly, there must be an odd number of arcs for a node if a traveler

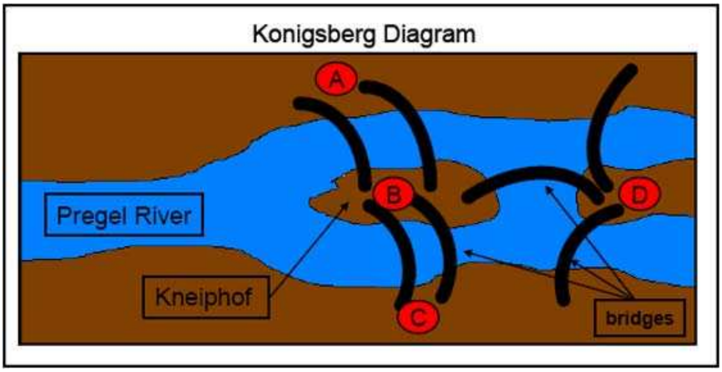


Figure 2. Konigsberg Bridge.

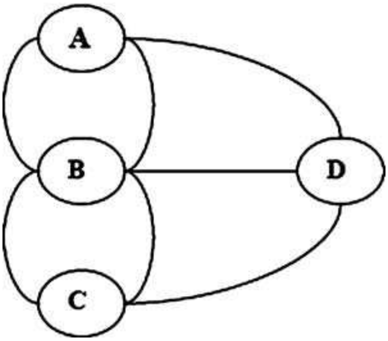


Figure 3. Abstract Representation of the Konigsberg Bridge Problem.

is to begin or end his path at that node, but if he begins and ends his path at the same node then there must be an even number of arcs at that node. So, for what networks does there exist a path such that one can visit each arc once and only once? For a solution to exist, there must be either zero (start and end at same node) or two (start and end at different nodes) nodes with an odd number of arcs and the rest of the nodes must have an even number of arcs. Any other graph has no solution to the question posed.

For the Konigsberg Bridge question, we can easily verify that each node has an odd number of arcs and since there are four nodes, there is no solution to the question. However, one may also easily verify that adding or subtracting any one bridge would make it a solvable problem. It just so happens that in 1875, another bridge was erected and it was indeed possible to visit each bridge once and only once. The appropriate path would be to begin and end at the two land masses not connected by this new bridge.

Another version of the question is to ask whether there exists a path beginning and ending at the same land mass, that will cross all seven bridges.

This version will also be answered in the same way, but solution is simply that a path exists only if there are no nodes with an odd number of arcs. If this is the case, one can start at any node and there is a path from that node that is a solution. Euler's solution gave birth to a new field of math known as graph theory, which would provide many rich methods used to analyze such problems. In addition, the field of network science eventually emerged from the field of graph theory.

2.4 Random Graphs

Paul Erdös and Alfréd Rényi made some of the first breakthroughs in network science. They modeled networks as random graphs, by first selecting a number, n, of nodes. They then connected every pair of nodes with a probability p, forming what came to be known as random graphs or what are sometimes called the "Poisson Random Graph" or "Bernoulli Graph."

Paul Erdös and Alfréd Rényi were the first to thoroughly study the structure of graphs. Mathematicians began to ask questions like, on a graph with n nodes where each arc had a probability p of being connected, what is the probability that the entire graph is connected? Where a connected graph is defined as one in which any node can be accessed by any other node. In other words, for all pairs of nodes, there exists a path between them. Certain properties of graphs may change as n, the number of nodes, increases and mathematicians often looked at the properties in the limit as n went to infinity.

More generally, it was found that random graphs often experience a phase change. A component of a graph is a set of nodes that are connected. The degree of a given node, is the number of arcs attached to it. This often is the same as the number of other nodes attached to it, but in some networks, there may be multiple arcs attaching two nodes. It was found that by varying the average degree of the nodes in a graph, the structure of the graph may change. Specifically, as the average degree increases, the random graph will experience a rather quick phase transition. Before this transition, the graph will be composed of many rather small components, the arc density will be very low, and the graph will not be connected. However, after the quick transition, the graph will have one "giant component" containing most nodes. The graph still may not be connected, but it will be close, in that most nodes will be connected to one another.

While it was found that these random graphs did exhibit some features such as the "small world effect" mentioned in the next section, these random graphs failed to display many of the other topological characteristics of real networks (to be discussed in the coming sections). Hence, they were left behind for newer models more capable of modeling the real networks around us. However, the approach of Erdös and Rényi to study topological properties of graphs would have lasting effects in network science because they built the foundation of how we fundamentally think about networks.

2.5 It's a Small World

Surely, at one time or another, everyone has been told "you are within six degrees of Kevin Bacon" [or something with essentially the same implication]. This phrase is one of the most popular examples of the small world effect that is within networks. The idea of the small-world effect was first published by Frigyes Karinthy in the short story Chains (Karinthy, 1929), where a character claimed that all people in the world were connected via at most five people.

The scientific history of the small world effect dates back to the research of Stanley Milgram (1967). Milgram randomly chose people in a seemingly far off place like Wichita, Kansas and in a second study, Omaha, Nebraska. He then sent them a letter with instructions on how to participate in his study. He told them to send it to a divinity student in Sharon, Massachusetts in the first study and to a Boston Stockbroker in the second study. However, there was a catch. The participants were only allowed to give the letter to a person they knew by their first name. The lists were mailed back to Milgram, so that he could trace the progress of the letters. Milgram found that the completed letters passed through an average of just under six people. While the methodology and results of Milgram's experiment have been contested by some, it is commonly believed that the social network that we all live in, can be characterized by a small world. At first glimpse it may seem surprising that random people chosen could be connected by an average of less than six degrees. However, if we consider how many people are within six degrees of us, then it may no longer be surprising. Each person easily has fifty friends connected to them, most people have hundreds.

Suppose we consider the random graph of Erdös and Rényi. Suppose each node has on average one hundred links. Thus a given node X, has one hundred other nodes within one degree of it, but each of these nodes also has an average of one hundred nodes linked to it. So there are 100^2 nodes within two degrees of X and 100^3 nodes with in three degrees and so on. Thus, there are 100^6 nodes within six degrees of X. This amounts to 1,000,000,000,000 nodes within six degrees of X. This surely cannot be an argument that identifies the number of people within six degrees of any person in the real world, since the current world population is only at about 6.5 billion. The error in thinking intrinsic to the argument just presented is fairly simple to find. Denote the group of nodes within one degree of X as X_1. Now each node in X_1 also has on average 100 nodes within one degree of it. However, some of those nodes within one degree of it are also in X_1. So there is a significant overlap in the links, causing there to be far fewer than 100^6 nodes within six degrees of X. This may be a bit

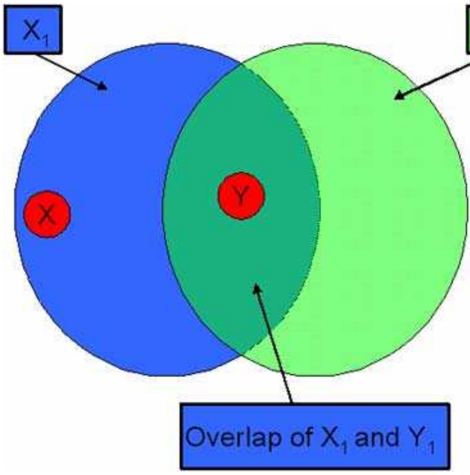


Figure 4. Overlapping Sets and the Small World Effect.

confusing, so let us denote node Y as a node in X and denote Y_1 as the set of nodes within one degree of Y. Now, some of the nodes in Y_1 may also be in X_1, as depicted in Figure 4.

This overlap in the links causes there to be far fewer than 100 nodes within six degrees of X. This may seem crippling to the small world effect, but it should be noticed that the current American Population is about 300,000,000 which is less than $100^{3.25}$ nodes. The degree of overlap between the nodes within one link of X_1 and those nodes in X_1 will surely affect the number of nodes within 6 degrees of X. However, the important thing to realize is that as the number of links L grows, the number of nodes connected to X within L links grows with a power relationship. The number of nodes within L links should be able to be modeled approximately as $\beta^{\alpha L}$where β is some way of estimating the average links per node and α is some parameter to estimate the amount of overlap in links between nodes.

This may make you think, do all nodes have the same number of links? If not, how could we estimate β? Is there a distribution for the number of links L that some node X has? These questions will be looked into a bit more when Hubs are considered. In addition, this model may bring new questions of how exactly do we describe or find α? This is one of the questions that could be answered by the clustering model presented next.

2.6 Clustering

The random graphs of Erdös and Rényi easily show the small world effect found in social networks by Milgram. However, network applications such as social networks often display other topological properties not captured by random graphs. One of the most obvious characteristics of social networks that random graphs cannot model is social cliques or what Watts and Strogatz call clustering (1998).

Random graphs connect any two nodes with a probability p. This means that if Adam is friends with Bob and Adam is also friends with Cathy, then Bob and Cathy have a probability p of being friends with each other (refer to Figure 5). David, not a friend of Adam, also has a probability p of being friends with Bob and a probability p of being friends with Cathy. So David is just as likely to be friends with Bob or Cathy as Cathy and Bob are of being friends with one another, even though Bob and Cathy have a mutual friend (Adam). Random graphs do not appropriately model social networks because this scenario is simply not realistic. Since Bob and Cathy are both friends with Adam, they are more likely to be friends with each other than with David, who is not friends with Adam.

Often people are a member of a social clique or group of friends where everyone is friends with everyone else or almost everyone else. Two people who are friends with Adam are more likely to know each other because they may meet at a common social gathering, including birthday parties, family events, work, or class. So this implies that if Adam is friends with Bob with a probability p and also with Cathy, with a probability p, then Bob and Cathy should be friends with a probability greater than p and David should be friends with Bob and Cathy, each with a probability lower than p. In Figure 5 below, this is shown by the thick solid lines representing the highest probability and the thin small dotted lines representing the lowest probability. Watts and Strogatz (1998) provided the first model that incorporated this property of clustering into the topology of the network, while still maintaining the small world effect (short path between any two nodes). They developed a model which began with the network in a ring, where each node was connected to the nodes close to it, but not those far away. In Figure 6, taken from Watts and Strogatz (1998), the network on the left has 20 nodes, each of which is connected to its 4 neighbors.

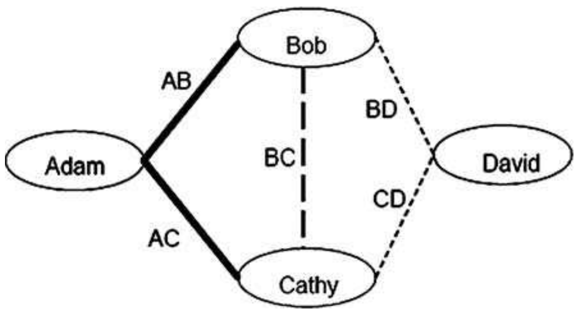


Figure 5. Small World Social Network.

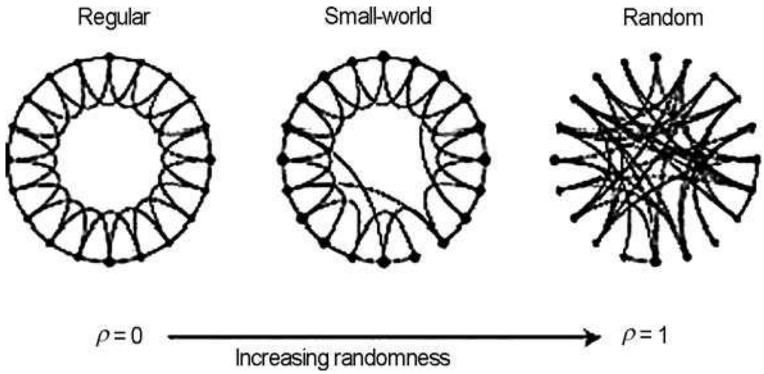


Figure 6. Randomness and Clustering.

Then the method proceeds by starting with a vertex and the edge that connects that vertex to the closest vertex clockwise to it. With a probability p we rewire it to another vertex chosen uniformly at random (without duplication allowed). Then we move to the next vertex in clockwise order. When the lap is completed we move on to the edge with the next shortest link without ever considering the same edge twice, again rewiring it with probability p. For small p Watts and Strogatz (1998) found that the small world shown in the middle diagram emerged. For this small world, the path length was rather short on average due to some arcs going across the circle, yet the clustering coefficient was very high because most neighbors were connected within one or two links. Since each neighbor is connected within one or two links, the neighbor of a node's neighbor is also very close allowing clustering to become a part of the topology of the network.

2.7 Hubs

Watts and Strogatz (1998) still did not fully capture all of the topological properties of many of the networks around us. In their model, each node begins with the same number of links. Even as p increases, since the links are rewired with a uniform distribution, the distribution of the number of links that a node has will not follow a power law distribution. Instead it will follow a peaked distribution. One may think this is not a significant issue, seeing as we should expect some type of bell shaped distribution from a random network. However, Albert-László Barabási has found that the distribution of links in various networks follow a power law distribution not a bell shape distribution.

Barabási found that various networks displayed a power law distribution. This means that the great majority of nodes have very few if any links, while just a few have the great majority of links. This seems intuitive when looking at networks around us. Google has millions of links while most personal web pages simply have one or two. In fact, the ten largest airports in the United States have flights to virtually any airport in the United States, if not the world. For example, a recent check revealed that Philadelphia International Airport had 255 flights arriving from 108 different airports. Yet, a local rural airport typically only offers flights to two or three other airports. In University Park, for example, the local airport has 8 flights arriving from 4 airports, all of which are major airports. There are about twenty to thirty very large airports (in the US) that most people fly out of when they need to travel a significant distance. Yet, there are over 19,000 airports in the United States.[1] From University Park, Pennsylvania, there are 10 airports within 110. Of these 10 airports, 9 are small and 1, Harrisburg, is a bit larger. But, even Harrisburg had only 42 flights coming in from 12 cities on a recent day when we checked the published schedules. When compared to Philadelphia, Harrisburg was linked to less than 1/6 the number of airports and had 1/9 the number of flights serving these other airports. Harrisburg is the largest of the 10 closest airports to University Park (within 110 miles) and Philadelphia is a large airport, but surely not the largest. In 2004, Philadelphia was the 16th largest airport in the United States. So for a network of airport connections in the U.S., there are more than 10 times the number of small nodes than there are large ones and the large nodes often have far more than 10 times the number of links. This means that the distribution of links of the networks around us surely are not bell shaped because of this small number of nodes that have often more than 80% of the total links incident upon them. Barabási refers, quite appropriately, to these popular nodes as "hubs". Barabási also began to call complex networks that displayed the power distribution scale free networks. Since there is no single node which can be chosen to characterize the population of nodes, there was no scale in these networks, and hence the term scale free network was coined.

2.8 Modeling Hubs

So if the Watts-Strogatz model cannot portray a power distribution of links, then how can we model it? Let's start by looking a bit deeper at the Watts-Strogatz model and try to find why it was bell-shaped to begin with. The model presented by Watts and Strogatz (1998) begins with each node having the same number of neighbors, denoted by k. Then through the rewiring process, p, the probability of an edge being rewired is the same for all nodes on the ring. Hence, when the rewiring is finished, each node has a minimum of half the number of arcs that it began with. In addition, since the arcs are rewired to a

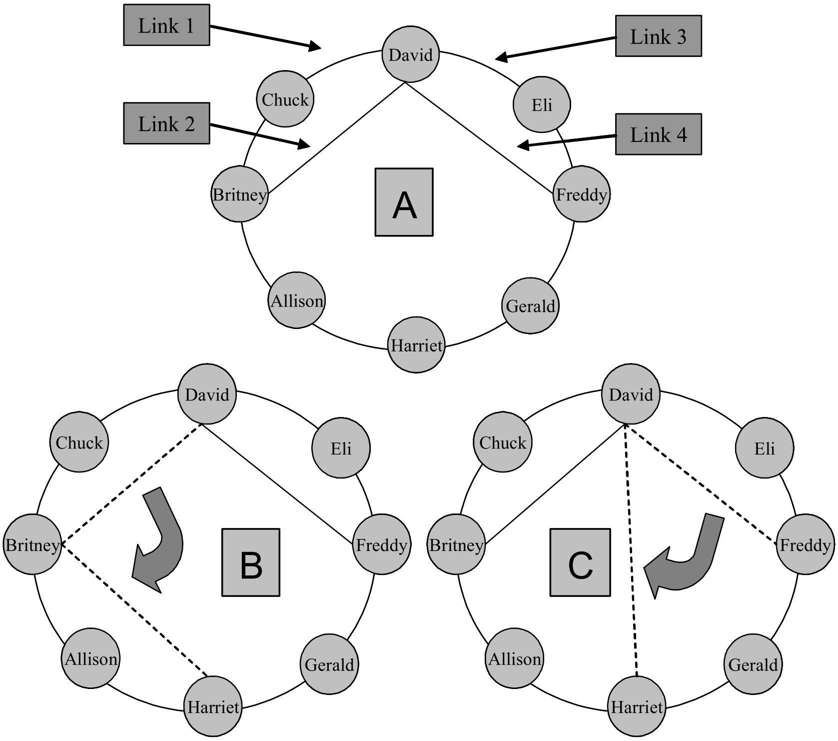
[1] In 2003, there were 19,581 US airports, but only 5,286 were public airports according to the Bureau of Transportation Statistics. Airline travel data can be obtained from the Bureau of Transportation Statistics (2006).

node with a uniform distribution, no node will gain that many more arcs than the rest.

The number of arcs gained by any node will have a normal distribution, while the number of arcs lost will have a binomial distribution. This will create a unique distribution for the number of arcs a node will have. More importantly, an upper and lower bound was just put on the number of arcs a node may have. The lower bound was created because each node must keep $k/2$ arcs. Since this may not be obvious, Figure 7 illustrates the process by which hubs are formed. Suppose we have a particular node David that is connected to two neighbors before it, Britney and Charles as well as two neighbors after it, Eli and Freddy (this makes $k = 4$). Now each of David's links to these other people may be rewired with probability p. Now, let's assume that in clockwise order, the order is Britney, Charles, David, Eli, and Freddy. Then we will randomly decide if Britney's link to David is rewired when it is Britney's turn, so this means if this link is rewired then Britney will be re-linked to someone else (chosen uniformly) and David will lose one of his four links (Britney is re-linked to Harriet). Similarly, Charles' link will go through the same process. However, when we consider Freddy's link to David, if it is rewired, then David will be re-linked to another node (chosen uniformly). David will obviously have to keep this link regardless of whether it is re-linked, while Freddy will be losing a link if it is re-linked. Figure 7 also shows David being re-linked to Harriet (chosen uniformly). In addition, Eli's link will go through the same process. Thus David must keep 2 links no matter what value of p or what random outcome occurs. In the general case, this is one half of the starting links or $k/2$.

In Figure 7, not only is there a lower bound on the number of links, but there is also an upper bound on the number of links a node can obtain. No node can obtain more than $N - 1$ links, where N is the number of nodes in a network. In a large complex network, this practically is not an upper bound. But there is a limiting upper bound, one that is more of a statistical upper bound. There are only $\frac{1}{2}NK$ links in the network and on average only $\frac{1}{2}pNK$ links will be rewired. Since there are N nodes in the network and each will receive a rewired link with equal probability, each node will receive rewired links according to a binomial distribution. Each node will keep its original links according to a binomial distribution as well. So, the number of links a node will have is the sum of two binomial distributions. This resulting distribution is bell-shaped and not a power distribution as the real complex network link distribution is.

At the core of his inquiries, Barabási (2002) asked two related questions: (i) why is the random graph model insufficient to model the power distributions of real networks? and (ii) how are real networks built? The first step in answering these questions is to recognize that a real network is not built by first making nodes and then randomly connecting them as Erdös and Rényi assumed, nor is a real network created by configuring nodes in a certain way



A) Original wiring for David in the Watts Strogatz Model
B) If the link between Britney and David is rewired (this occurs with probability *p*) so that Britney is linked with Harriet (Harriet is chosen uniformly). Now David lost a link, he would have 3 not 4.
C) If the link between David and Freddy is rewired (this occurs with probability *p*) so that David is linked with Harriet (Harriet is chosen uniformly. Now David keeps the link and Freddy is the one who lost a link.
Note also that Link 1 and Link 2 will both be lost by David, if rewired (with probability *p*) and Link 3 and Link 4 will both be kept by David, even if rewired. Thus David, must have a minimum of 2 Links in the end.

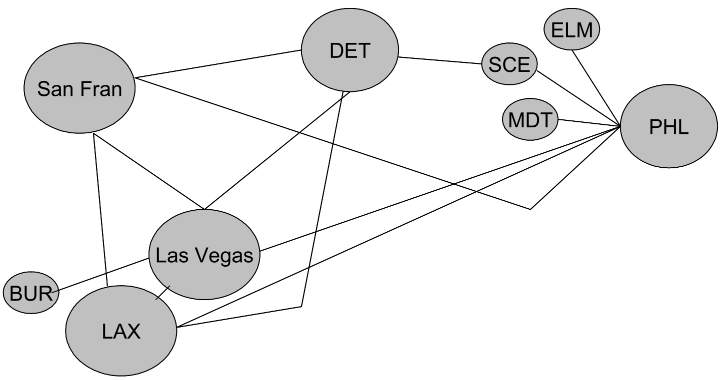
Figure 7. Emergence of Hubs.

and then changing the links as Watts and Strogatz (1998) assumed. When a company, say an airline decides which cities to link with a flight, do they flip a coin to decide whether to offer a certain flight? When David is trying to decide who will be in his social circle, does he flip a coin to decide whether he should ditch one friend for another? Does the airline company flip a coin to decide whether to cancel the Los Angeles to Philadelphia flight and replace it with a Philadelphia to State College flight?

The truth of the matter is that most real networks are formed by just a few nodes and a few links. The original links may be random or may not, but – as the network grows – more and more links are added. However, they are not added completely randomly or at least not uniformly so. Barabási hypothesized that they are added with preferential attachment. That is, each new link would be more likely to be added to those nodes which already had more links. Thus, the bigger cities would be more likely to have new flights added. This allows the rich to get richer, so to speak, and the big nodes to get huge, while the small ones stay small. The combination of allowing a network to grow and allowing it to grow with preferential attachment, forms the power distribution characteristic of scale free networks.

2.9 Vulnerability and Epidemiology

As we have seen, the model of network evolution proposed by Barabási often favors older more established nodes, but it does not preclude younger ones. When a node is first created, it will only have a few links and those links will be more likely to be connected to older more popular nodes. This means that many of the smaller nodes are linked to the popular nodes, that Barabási calls hubs. These hubs allow the existence of the small world effect and clustering. This can be seen in the airline example of Figure 8, fairly easily. The University Park airport is a very small node, connected to only four other airports.



University Park, PA (SCE) is connected to:

1. Philadelphia, PA (PHL) and Detroit, MI (DET) within one link
2. Elmira, NY (ELM) and Harriburg, PA (MDT) within two links
3. Las Vegas, Nevada and Los Angeles, CA (LAX) within two links
4. Burbank, Ca (BUR) within three links

Figure 8. Airline Network Connecting University Park to the World.

Recently the largest of these four was Detroit (9th) and the smallest was Dulles in Washington, DC (23rd).[2] So basically, this very small node was connected to four hubs. These hubs are surely connected to all the other hubs, even on the other side of the county, like San Francisco or Dallas. So, University Park is connected within one link to 4 hubs, two in the Midwest and two in the East. From these hubs, University Park is connected to all of the hubs in the United States. So University Park is within 2 links of any hub in the country. Then each small airport in the country is connected to a hub, so University Park, is connected to virtually any other commercial airport in the US by only 3 links. This results in quite a small world, from the perspective of airport connectivity. In addition, of the 10 airports within 110 miles of University Park, most of them will have flights to Philadelphia as well; making them within 2 links of University Park and also causing the topological property we earlier called clustering.

It is precisely the emergence of hubs that allows the small world effect to occur. This topology induces a few other properties as well – including quick diffusion of information and innovations as well as increased vulnerability. In particular, Thadakamalla et al. (2004) have investigated what they called the survivability of a network. According to Thadakamalla et al. there are four aspects of survivability. The first of these is called the robustness of a network: robustness can be measured by measuring the connectedness or the size of the largest component, after a number of nodes are removed. A robust network should be able to survive both random node breakdowns as well as targeted attacks. Some networks will breakdown to various components that cannot communicate with one another, after only a few nodes breakdown. A second aspect is the responsiveness of a network, which is a measure of how quickly the network can be traveled through. It is usually measured by averaging the shortest paths between all pairs of nodes. A third aspect, flexibility of a network, measures how often there exists an alternate path between any two nodes. The clustering coefficient discussed earlier, is a good measure of the flexibility of a network. The fourth aspect, the adaptivity of a network, is a measure of how easily paths can be created and destroyed in order to change a topological characteristic of the network.

Each of the four network characteristics identified by Thadakamalla et al. is important in measuring how well a network will perform in the event that nodes or arcs breakdown. In fact Thadakamalla et al. ran simulations on several networks showing that scale-free networks were quite robust when random nodes broke down. This means that quite a few nodes can fail and yet the rest will still be connected to each other. Yet, scale-free networks are quite vulnerable to attacks on their hubs. By contrast, random networks and small

[2] Airline Statistics from the Bureau of Transportation Statistics (2006).

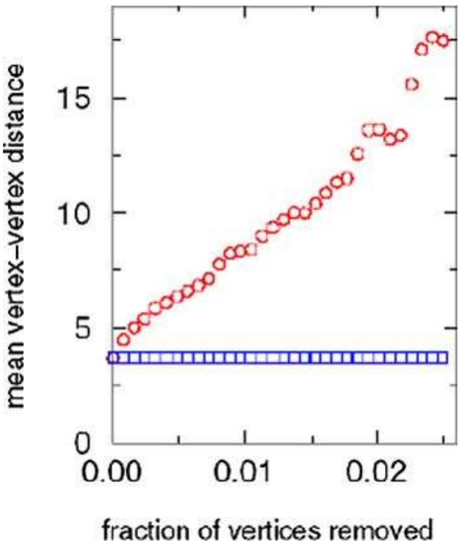


Figure 9. The Effect of Vertex Removal.

world networks are less vulnerable to attacks, but more vulnerable to random breakdowns. Albert, Jeong, and Barabási (2000) studied the vulnerability of scale free networks. They did this by measuring the average vertex to vertex distance as vertices are removed by random failures and by targeted attacks. As seen in Figure 9 taken from Newman (2003), the mean vertex-to-vertex distance is hardly affected by random attacks (squares), yet the mean vertex-to-vertex distance significantly increases for targeted attacks (circles). Thus, we can see that scale free networks survive random failures quite well, but do not easily survive targeted attacks. These two studies used different measurements, but in fact they show similar results.

The vulnerability of supply chain networks and the stubbornness of some epidemics can be analyzed in a similar way with network structure. In particular, the SIR (susceptible/infective/removed) model is one mathematical model of epidemics. It breaks up the total population into three sets of people: those who are susceptible (S), infected (I), and removed (R). In the SIR model, the population of each group is governed by coupled differentials equations relating the populations. The basic SIR model given by Newman (2002) is the system of equations in (1). Most mathematical models of epidemics, including the SIR model assume that each person is equally likely to come into contact with an infected person. However, this is simply not the case in most real networks; so a rather challenging task arises – namely that of imposing a hub-based topology on the simple set of ordinary differential equations comprising the above model. Thus, the mathematical theory of epidemiology – as well as that of population migration and innovation diffusion – may need to be rethought.

$$\frac{ds}{dt} = -\beta is \qquad (1)$$
$$\frac{ds}{dt} = \beta is - \gamma i$$
$$\frac{dr}{dt} = \gamma i$$

The above challenges not withstanding, it is constructive to consider what might be the insights on epidemic management afforded by network science. By turning the vulnerability of scale free networks on its head, it can be seen that if a social network has a scale-free nature then the spread of a virus can be stopped by targeting key people. If we vaccinate certain key people in a social network, it is the same as making a targeted attack to break down the connectivity of the social network. However, since random attacks on a scale-free network, make a negligible impact on connectivity, this implies that random vaccinations on a social network will make a negligible impact on stopping the spread of a virus. The topology of a network can grossly effect the spread of an outbreak over a network. Watts (2002) discusses cases in which for one type of topological structure an outbreak may simply die out, yet in another it will spread to every site (epidemic), while in other topological structures it will depend on how quickly the infection can spread. This is one way in which topological properties such as degree distribution and clustering can effect the spread of disease.

Similarly, an epidemic's transmission network may be compartmentalized in such a way that an infection cannot spread quickly. Often the term community structure is used to describe the property of networks where the nodes may be split into different "classes" which have higher clustering coefficients within each class.The spread of a Sexually Transmitted Disease will depend on a social network. However, most people are more likely to interact sexually with others of similar age and financial status, the same ethnicity, and different sex. So, people do not simply interact uniformly with all other people allowing the structure to play an integral role in the spread of an STD. For this reason, since the SIR model assumes "full-mixing", it is not always a good approximation. Often in social networks, an epidemic may spread more slowly.

Power distribution networks and communication networks are similar to epidemics viewed as networks, in that it is important for the network not to breakdown if only one node or arc suffers failure or interdiction. However, in these networks when an arc or node breaks down, the remaining load is redistributed to the rest of the network. This may result in overloading more nodes or arcs, which then breakdown and the load is redistributed again. Hence there are cascading failures until the network stabilizes. Boccaletti et al. (2006) discuss the effects that topological properties, such as path redundancy, load and capacity, can have on cascading network failures.

When it comes to viruses spreading through computer networks, however, Balthrop et al. (2004) argue that this targeted vaccination technique may not work as well. Viruses can "choose" a network topology by choosing their mode of transmission: email, IP addresses, etc. Some of these networks, such as the network created by IP addresses, are quite uniform while others such as the one created by email traffic are close to scale-free. A virus can spread through any of various networks created via a mode of transmission. So, if a virus is attacked by vaccinating the 10 percent of the population most at risk, little may be achieved if there is unknown scale-freeness. Balthrop et al. argue that any vaccination technique that requires knowledge of the topology or assumes a certain topology will be ineffective because networks topologies are constantly changing and the virus may change the network it is acting on by changing its method of infection.

Thus, Balthrop et al. (2004) propose instead a method called "throttling" to thwart the spread of computer viruses. Many viruses depend on contacting other machines hundreds of times a second in order to cause an epidemic. However, people can at most contact a few computers a second and usually much less. Throttling proposes to take advantage of this disparity by slowing down the contact rate of computers. By only letting a computer contact one other computer per second, people's legitimate information flow is not slowed down because this does not limit most people. However, a virus which may need to make hundreds of contacts per second, will be slowed down by a factor of 100. This means, however, that the epidemic propagates much more slowly. With "new" time created in this fashion, traditional anti-virus software can be updated to destroy a new virus. If throttling can slow a virus down to doubling once per day or week, ordinary users can with – a near certainty equivalent – update their antivirus software and stop the epidemic. Throttling in theory will work on any network, regardless of topology.

2.10 Network Games

The topology of complex networks has been discussed in prior sections, but now I turn to look at the routing of flows over a given network. The most common example, is the traffic network that most of us travel through each day on the way to work. Each person travels from their own home to their workplace along a route that they expect to minimize their travel time. The problem that each of us has surely encountered, is that of traffic or congestion. A road can only hold so many people and as the number of people on a road increases, the time it takes to travel across it increases in some way also. Additionally, each person does not consider the fact that for each road they choose to take on the way to work, they are increasing the travel time for every other person who needs to take that road. That is, they act selfishly in noncooperation with others.

This network can be modeled in game theory as a complex network of nodes and arcs, with infinite agents. Each agent has its own source and destination (not necessarily unique) and controls a small fraction of the flow over the network. Further, we assume that each person acts in a selfish manner, that is, they do not consider the effect on others, when selecting their route. They simply try to minimize their travel time. The cost that each agent incurs over an arc is called the latency and the total flow over an arc is called the load. It is generally assumed that the latency is a nondecreasing function of the load.

The Nash Equilibrium is the set of routing or set of flows that is reached when each person travels selfishly. For the Nash Equilibrium, each agent will incur the same cost for a given source and destination. That is, any two agents that have the source destination pair, will incur the same latency regardless of the route they took. This may not be immediately intuitive. To clear things up, let us ponder the case if it were not true. Suppose agent A incurs a latency of x and agent B incurs a latency of y for the same route and $x < y$. Then soon enough, agent B would discover agent A's route that is less costly and would then switch to agent A's route. But, then the original solution would not be at equilibrium. Thus, in order for an equilibrium to occur, all agents must incur the same cost for a given source and destination.

Since each person chooses their route in such a way as to minimize their own travel time, the total travel time traveled by all (Social Cost) is higher than it would be if this were minimized in a regulated network. Now, let me call the socially optimal solution that which is achieved if the social cost is minimized in a regulated network. This leads to the question, "How much worse is the Nash Equilibrium then the social optimal?" The difference between the Nash Equilibrium and the Social Optimal is commonly referred to as the "Price of Anarchy."

Roughgarden (2002) gives a simple example created by Pigou (1920) and given in Figure 10, to display the difference between the Nash Equilibrium and Social Optimal. Pigou's example consists of one unit of flow that has to travel

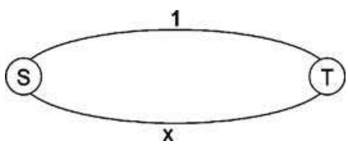


Figure 10. Pigou's Example.

from s to t. The cost or latency of the top arc is 1 regardless of the load over the arc, whereas the bottom arc has a latency equal to the flow over it. If there is more than one unit of flow over the network, we can think of x as the fraction of the total flow.

Now the Nash Equilibrium consists of the full unit of flow going on the bottom arc, that is $x = 1$. Why? Well suppose it were just a bit less, so $x < 1$. Then this implies that some agents are using the top arc and incurring a cost of 1 unit while others are going on the bottom arc and incurring a cost less than 1. So, clearly those agents incurring the cost of 1 on the top arc, would switch to the bottom to save a bit. This would continue, until the cost of the top arc equals that of the bottom, which happens to occur in this example when there are no more agents using the top arc and all are on the bottom.

Yet this Nash Equilibrium is clearly not the social optimal. The social cost is simply the total cost incurred which equals:

$$SC = f_{top} \cdot l_{top}(f_{top}) + f_{bottom} \cdot l_{bottom}(f_{bottom})$$

where f is the flow over the arc and l is the load dependent latency over the arc. So

$$l_{top}(f_{top}) = 1$$
$$l_{bottom}(f_{bottom}) = f_{bottom}$$

lets denote f_{bottom} as x so

$$\begin{aligned}
f_{bottom} &= x \\
f_{top} &= 1 - x \\
l_{bottom}(f_{bottom}) &= l_{bottom}(x) = x \\
l_{top}(f_{top}) &= l_{top}(1-x) = 1 \\
\Longrightarrow SC &= (1-x) \cdot (1) + (x) \cdot (x)
\end{aligned}$$

then to find the x that minimizes SC is elementary calculus:

$$\begin{aligned}
SC &= x^2 - x + 1 \\
\frac{\partial SC}{\partial x} &= 2x - 1 \\
\frac{\partial SC}{\partial x} &= 0 \Longrightarrow 2x - 1 = 0 \\
\Longrightarrow x^* &= \frac{1}{2} \\
\Longrightarrow SC^* &= \left(\frac{1}{2}\right)^2 - \frac{1}{2} + 1 = \frac{3}{4}
\end{aligned}$$

So, the social cost is minimized with a flow of $\frac{1}{2}$ on each arc giving a Social Cost of $\frac{3}{4}$. Note that this is different from the Nash Equilibrium, which had all of the flow on the bottom arc ($x = 1$), but had equal costs for each arc ($l_{bottom}(f_{bottom}) = l_{top}(f_{top}) = 1$). The arc latencies for the Social Optimal are 1 for the top and $\frac{1}{2}$ for the bottom. So those who take the top arc spend just as much time, while those on the bottom save. Indeed, selfish routing does not gain anything for anybody but simply results in some people being more hurt than they need be if the socially optimal solution were used.

Tim Roughgarden and Eva Tardos (2002) found bounds for the price of anarchy. They found when the arc latencies are a linear function of the arc flows, the total latency of flows for the Nash Equilibrium will be no more than $\frac{4}{3}$ of the total latency incurred by the optimal regulated routing. However, when the latencies are not linear, but instead simply monotonic with respect to flows, the price of anarchy can still be bounded. Now the flow of selfish routing is bounded by the total latency achieved by routing twice as many units though the regulated network.

For any Nash Equilibrium solution of flows in a network, each path from a particular source to a particular destination must have the same total latency (commodity). If this were not true, then there would exist an alternate route for the same source and destination, with a lower latency. But, if this were the case and each agent acts selfishly, they would surely prefer to switch to such a lower latency route implying the non-optimality of such a solution. Beckman et al. (1956) first showed such properties of the Nash Equilibrium Solution.

Braess's Paradox has captured much of the work done in the game theoretic applications in networks. The basic idea of Braess's paradox is that for a given network it is possible that adding an additional arc may result in increasing the total latency as well as the latency of each agent for the Nash Equilibrium Solution. At first blush, one may wonder how this could be true? Hence why it is called a paradox. Roughgarden (2002) gave the example in Figure 11, which should help demystify the paradox:

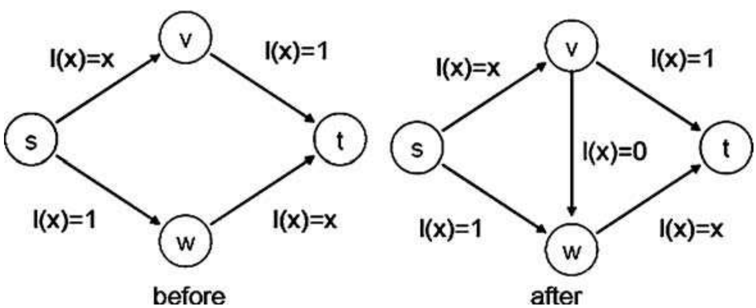


Figure 11. Braess's Paradox.

As Roughgarden explains, in the first network (before), the Nash Equilibrium for one unit of flow is for half to go on the top route and half on the bottom, each paying a total latency of $\frac{3}{2}$. In this example, the solution also happens to be the socially optimal solution as well. Now, if another arc with 0 latency is added, the Nash equilibrium will change. In the new network (after), the Nash Equilibrium solution will be for the entire unit of flow to take the path s, v, w, t with a total latency of 2, which is greater than before. Meanwhile, the socially optimal solution stays the same. This again, may not be intuitive, but suppose a small fraction or a single agent decides to deviate from this path (s, v, w, t), then it must take either (s, v, t) or (s, w, t). Now, if the agent decides to instead take (s, v, t), then it will travel along (s, v) with all the other agents, saving nothing. However, it will then have to travel across (v, t) for a cost of 1 while all the others travel (v, w) for 0 and then (w, t) for less than 1 (since this agent just left this path). Seeing that all the agents traveling (s, v, w, t) have a lower total latency, this agent should not make such a change in path. A similar argument can be used to show that (s, w, t) is also not a good choice. So, the Nash Equilibrium results in each agent having a higher total latency in this example, but how much higher?

Suppose we route two units through the (after) network, with half (1 unit) taking (s, v, t) and half (1 unit) taking (s, w, t). Then the total latency is 4. According to Roughgarden, in the general case, the flow of selfish routing would be bounded by 4. Since, the latency functions are linear though, the total latency of selfish routing is bounded by $\frac{4}{3}$ that of the social optimal of $\frac{3}{2}$, which is 2. Thus, the selfish routing in this network is bounded by 2, which is exactly what the Nash Equilibrium total latency is, so it is a case of the worst case scenario.

The cases considered by Roughgarden assume that the latencies of each user are independent of each other, that is the latency functions are separable. Perakis found bounds for the considerably more complex case of non-separable latency function Perakis (2004). Refer to Perakis (2004) to read more about these more general cases. This is simply one extension of Braess's paradox, there has been a host of literature published in recent years on various aspects of this topic.

It is important to keep the results of Braess's paradox in mind because sometimes it may be tempting to think that adding arcs or loads to a network will increase flexibility or clustering coefficients or some other property, making the network perform better, but while one characteristic may have been made better it may be adversely effecting the performance of the network. For very large networks, such adverse effects are difficult to find and analyze, so this presents a potential danger in many large networks in communication, supply chain, power supply, traffic, and others.

3. Nonlinear Dynamics

Chaos is defined by Merriam-Webster's Dictionary as "the inherent unpredictability in the behavior of a natural system." However, chaos is often misunderstood in its everyday use. Often, people mistake the fact that chaos is unpredictable for the idea that it is random. In fact, a chaotic system is deterministic, yet unpredictable. A chaotic system is one in which a very small difference in an initial condition, so small it may not be able to be measured accurately, can result in significantly different results. So, the system is unpredictable, but is not random.

Chaos is just one section of a broader field known as dynamics or nonlinear dynamics. Dynamical systems are systems governed by deterministic laws, however, sometimes even deterministic laws can result in unpredictability. The discovery of these unpredictable, chaotic systems raised the interest in the field of Dynamics.

3.1 A Brief History of Dynamics

In 1887, Henri Poincaré entered a contest on which he was supposed to show that the solar system was dynamically stable according to Newton's Mechanics. Although, he could not do so, his work was revolutionary and the judges (including Weierstrass) awarded him the prize anyway. As it turns out, instead he was the first person to stumble upon a chaotic dynamical system. Poincaré argued that even if the equations governing the system are known and deterministic that "small differences in the initial conditions produce very great ones in the final phenomena. A small error in the former will produce an enormous error in the latter. Prediction becomes impossible..." (Nijkamp and Reggiani, 1992).

The basic idea is that systems exist such that a small perturbation in initial conditions can cause differences later that are not only numerically significant, but in some applications, qualitatively significant as well. In such a system, the two paths are obviously quite close at the initial time and thus it is easily conceivable that they can again come close and even look quite similar for long lengths of time. However, they are not the same and they lead to quite different results. So, no matter how precisely such a system is measured, within that precision, the error in measurement could result in quite different states at a later time. This makes it impossible to predict the future state of a system, even if the laws governing it are perfectly known and perfectly deterministic. This lack of ability to predict the future state is seemingly random, but to call it such would be a mistake because the system is governed by deterministic laws. Systems such as these that are governed by deterministic laws are known as dynamical systems and chaos is simply one class of dynamical systems.

Unfortunately, although the work of Poincaré was immediately recognized as valuable, like many other discoveries, it was neglected for many years. This was most likely because such chaotic effects took place in infinite iterations and empirical experimentation was not practical for the time. His foresight would have to wait until the invention of the computer to be recognized for its full value.

In 1961, a meteorologist Edward Lorenz created a simulation model of weather conditions. One time he wanted to see the simulation a second time, so he attempted to rerun his model from a point somewhere in the middle, instead of the original conditions. To do this, he took the computer printout of the state conditions at the new time he wanted to start at and plugged them into his computer. To his surprise, the results were drastically different from the original results that he wanted to repeat. Figure 12 from Stewart (1989) shows how the two trajectories diverge.

His model was composed of deterministic equations. How could the results be so different the second time? He entered the data that he had from the same path and the equations were deterministic and yet the results were different.

He was eventually able to track it back to the fact that the computer used six-digit numbers and his computer printout cut off the numbers with three digits. So when he reentered the data for the second simulation, he only entered three digits, but when the computer ran the simulation the first time, it had all six. The conventional wisdom would tell most people that the last three digits are not very significant. In most calculations, it would be considered accurate to have three digits and six is above and beyond necessary. Yet, in this case the perturbation caused by the fourth digit at some time in the middle of the simulation was able to cause drastic changes in the results.

Figure 12. Lorenz Simulation.

3.2 The Basics of Dynamics

Dynamical systems are systems that have a state (represented as a variable or vector of variables) that changes over time in a manner which is governed by deterministic laws or equations. That is, the system is dynamic in a deterministic sense. Below are two examples:

$$x_{t+1} = f(x_t) \tag{2}$$

$$\dot{x} = f_1(x, y) \tag{3}$$

$$\dot{y} = f_2(x, y)$$

The dynamics are often given as a system of recurrence relations (as in 2) or differential equations (as in 3). The recurrence relation can be iterated infinite times to find the trajectory of the system as shown in Figure 13. A similar trajectory can be found for the differential equations by using small discrete steps to approximate a solution.

To "solve" the system, an equation must be found to directly find the state at any time in the future (find an equation $x(t)$ where x is the state of the system). Dynamical systems can be linear or nonlinear, that is the functions f, f_1, and f_2 in (2) and (3) may be linear or nonlinear functions. Linear systems have the property of being able to be easily solved, yet they often do not provide the array of behavior that can be found in nonlinear systems.

The dynamics of a system often involve parameters, which may or may not change over time, but do effect the nature of the dynamics either way. Often these parameters are responsible for the character of the system, causing it to be stable, unstable, or even chaotic for different values of the parameters. For example, the system shown above in (3) could have a solution for x and y

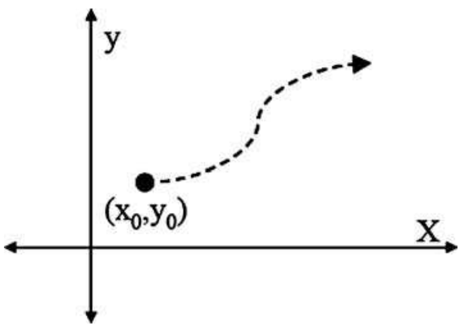


Figure 13. Trajectory.

(trajectory whose x and y components are determined parametrically by these equations):

$$x(t) = x_0 e^{at}$$
$$y(t) = y_0 e^{bt}$$

Now, the values of a and b will greatly affect the trajectories of the system. Think about what may happen if $a > 0$ as opposed to $a < 0$ and similarly for b.

3.3 Phase Diagram or Phase Space

In Dynamics, it is often easier to see the character of a system by looking at the phase diagram or phase space. In a phase diagram, each state of a system is plotted by its dimensions (so time is not included), so it is a simple point in the plot, sometimes called a phase point. Each state is then connected to the state preceding it and succeeding it. For example, a system with two variables x and y, would have a phase diagram with one axis for x and one axis for y. Then each state $S_1, S_2, S_3, \ldots$ are plotted as $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots$, where x_i is the value of x at time i and y_i is the value of y at time i. For a discrete time dynamical system we must then connect the points $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots$ to see the trajectory, but in a continuous dynamic system the transition between states are continues curves in the plane. The trajectory of Figure 13 is an example of a continuous time phase diagram. In a discrete system, only the points are plotted and the line connecting them is only to visualize a path between the states as shown in Figure 14.

It should also be noted that the particular trajectory that a system follows is dependent upon the initial starting point. That is, the trajectory may be significantly different if it begins from a different starting point.

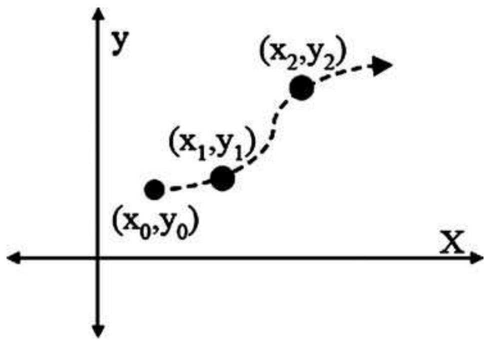


Figure 14. Discrete Trajectory.

3.4 Fixed Points or Equilibrium Points

Given the system

$$\dot{x} = f_1(x, y)$$
$$\dot{y} = f_2(x, y)$$

we can see that at any given point (x, y) we can calculate and find $\dot{x}$ and $\dot{y}$. In fact, $v = (\dot{x}, \dot{y})$ is the velocity vector for any given point (x, y). The vectory field V of velocity vectors shows the flow of the system. Those points where $v = (\dot{x}, \dot{y}) = (0, 0)$ have no flow and are thus called fixed points or equilibrium points.

There are two types of fixed points, those that are stable and those that are unstable. The stable points are fixed points whose neighboring points have velocity vectors pointing towards them, showing that the flow moves into them, and thus they are often called sinks or attractors. Similarly, the unstable fixed points have neighboring points whose velocity vectors point away from them, showing that the flow moves away from it, giving them the name repellers or sources. Take the system

$$\dot{x} = x$$
$$\dot{y} = y$$

whose vector field is shown in Figure 15.

At (0, 0), $\dot{x}= 0$ and $\dot{y}= 0$, which means that the system will not move from $(0, 0)$. Hence it is a fixed point. As we can see all of the vectors point away from the fixed point $(0, 0)$ meaning that the flow of the phase point beginning near $(0, 0)$ will flow away from it. Thus, this point is unstable and is often referred to as a source or repeller. A very similar system is:

$$\dot{x} = -x$$
$$\dot{y} = -y$$

whose vectory field is shown in Figure 16.

Again at $(0, 0)$, $\dot{x}= 0$ and $\dot{y}= 0$, so $(0, 0)$ is again a fixed point. As we can see all of the vectors point toward the fixed point $(0, 0)$, meaning that the flow of the phase point beginning near $(0, 0)$ will flow toward it. Thus, this point is stable and is often referred to as a sink or attractor.

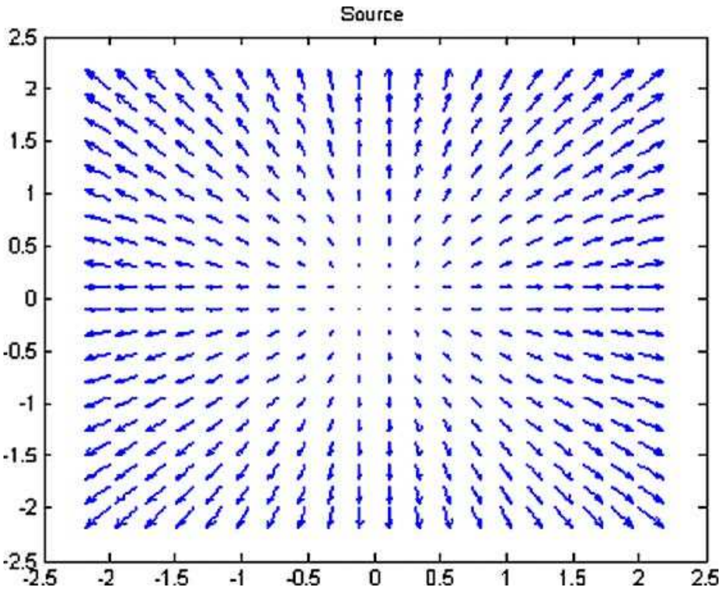


Figure 15. Source.

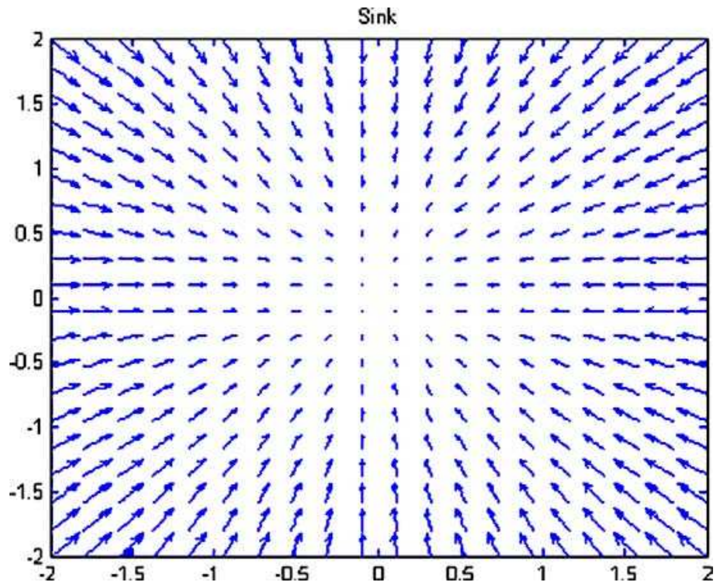


Figure 16. Sink.

There is yet a third type of equilibrium point, called a saddle point. A saddle point has two axes or manifolds going through it. The trajectories

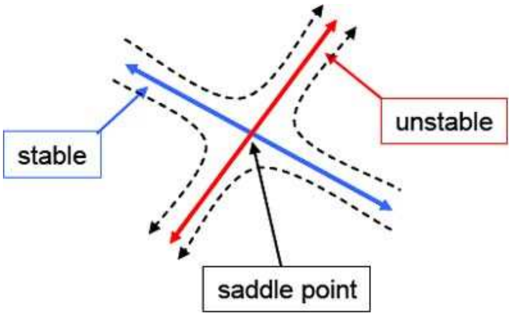


Figure 17. Saddle Point.

are asymptotic to one of the axes, known as the unstable manifold (refer to Figure 17) in forward time. The other axis is the stable manifold, the trajectories asymptotically come from this axes, that is in reverse time or the reverse trajectory is asymptotic to the stable manifold. At first blush it may seem odd that the trajectories approach the unstable manifold. So, it may help to think of it this way: in the direction of the stable manifold, the trajectories approach the saddle point as time goes to infinity, making it stable in this direction. However, in the direction of the unstable manifold, the trajectories go away from the saddle point, they are infinitely far away as time goes to infinity. It should now seem much more reasonable to label the stable manifold as stable and the unstable manifold as unstable.

The examples of systems in Figure 17, can once more be slightly modified to:

$$\dot{x} = -x$$

$$\dot{y} = y$$

whose vectory field is shown in Figure 18.

Now we can see that the vector field forms a saddle point at $(0,0)$ and the x-axis is the stable manifold and the y-axis is the unstable manifold.

An important note to make is that in the above three systems I have simply added a negative sign in one or two places and the topology of the vector field has changed significantly. So, now we can see that if we plot the vector field for the system and then we are given a starting position, we can follow the vectors to form a trajectory from that starting position. This type of analysis does not involve time, so we do not know how long it will take to follow a trajectory, but we know it will eventually follow the trajectory. The important think to realize is that minor changes in the differential equation governing the dynamics can significantly alter the topology of the vector field and thus the solution paths.

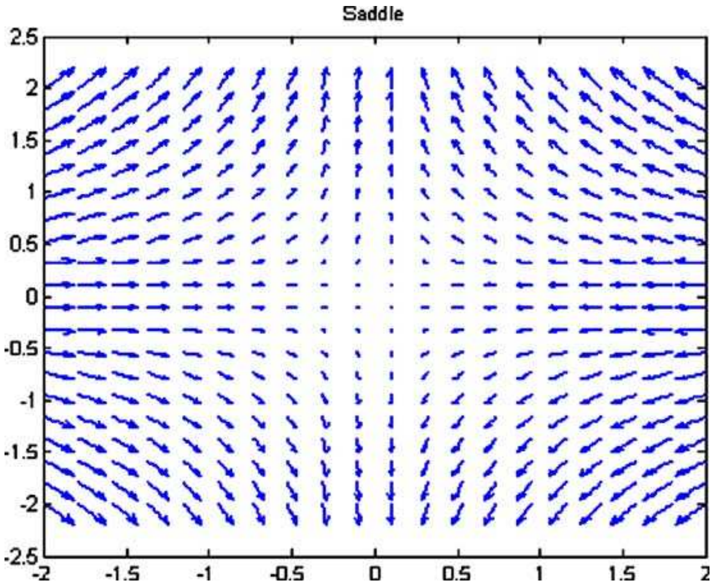


Figure 18. Saddle Vector Field.

3.5 Closed Orbits and Oscillators

A closed orbit is a loop or orbit, in which a system will periodically cycle through. That is, if the system starts on the cycle, then it will continue to go around it infinitely many times. At any time, it will visit each state infinitely many more times in the future. An example is:

$$\dot{x} = -y$$

$$\dot{y} = x$$

whose trajectory is shown in Figure 19.

The vector field for the system is shown in Figure 20.

3.6 Attractors and Repellers

By plotting the trajectory from different initial points, one can then see the behavior of the system. Many systems are periodic and will have a cycle or orbit in the phase diagram, no matter where they begin from. Others will have an asymptotic behavior where they will be attracted to a point or line. Still other systems may show a spiral towards a single point in the center, like a hurricane or water flushing down a toilet. Systems may show that different trajectories are attracted to an orbit from nearby locations. The term attractor refers precisely to these occurrences. The attractor may be a point, loop, or multidimensional

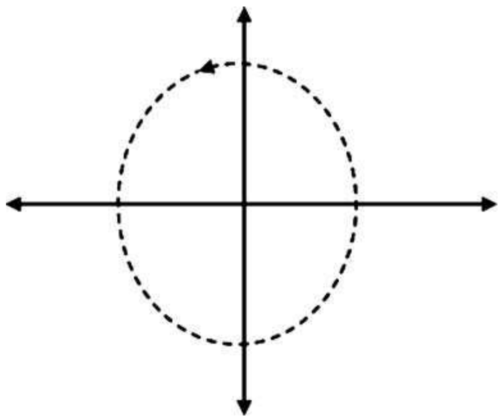


Figure 19. Center.

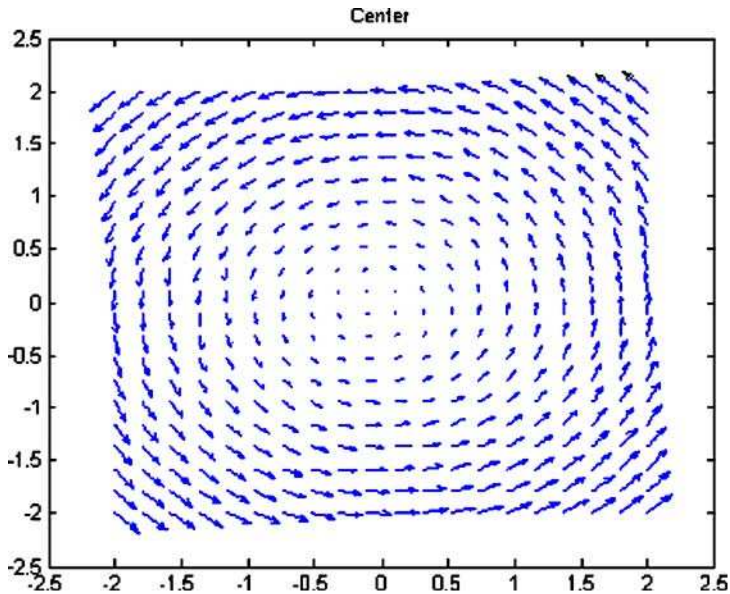


Figure 20. Center Vector Field.

loop that the trajectory moves towards as time or iterations progress. In a more mathematical view, the attractors are points or sets of points that the system approaches in infinite time.

A given fixed point, $\bar{x}$, is a local attractor if it is approached from starting points, x_0, that are within a certain neighborhood of $\bar{x}$. Similarly, $\bar{x}$ is a

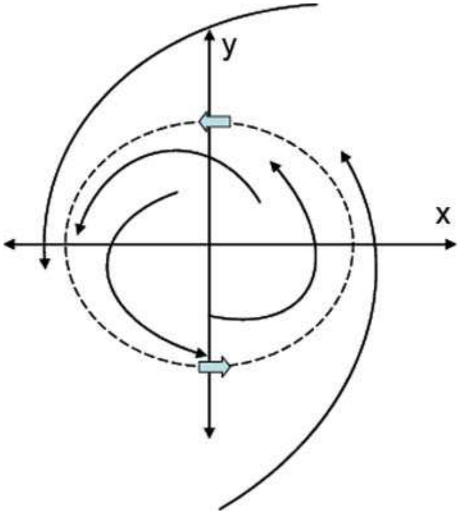


Figure 21. Limit Cycle.

globally attracting fixed point, if it is approached in infinite time by all starting points, x_0.[3]

There are various types of attractors and repellers. A fixed point is a single phase point which can either attract or repel trajectories (e.g. a stable fixed point, unstable fixed point, saddle point). There may also be attractors and repellers that are limit cycles. A limit cycle is a closed orbit, that is surrounded by trajectories that are not closed. Figure 21 is an example where the dotted trajectory is closed, but all the trajectories on the inside spiral toward it and all of the trajectories on the outside spiral toward it as well. This is called an attracting limit cycle. Another type of limit cycle will have a closed orbit, but the trajectories are repelled from it, such a limit cycle is a repelling limit cycle. Another case in where the system has a closed orbit with trajectories on the inside that are attracted, yet on the outside they are repelled or vice versa. This is neither attracting nor repelling in the sense described above.

Alternatively, in the case of a purely oscillating system, the phase portrait will consist of infinitely many concentric loops shown in Figure 22. The vector field for such a case is shown in the vector field for the center shown in Figure 20.

There are several types of attractors and repellers in two dimensions including fixed points and centers. Additionally, systems in three dimensions will increase the variety of attractors and repellers to include what are called

[3] The above definition of attractor is based on the definition of assymptotic stability. Liapunov Stability is another type of stability.

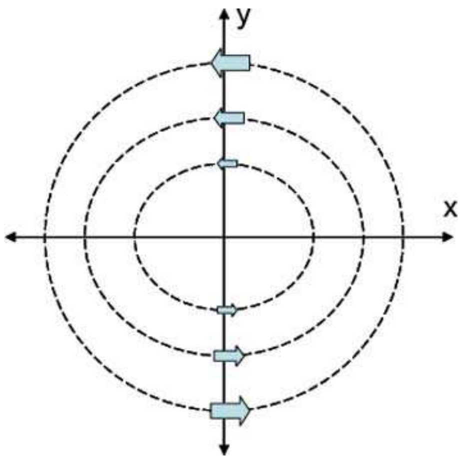


Figure 22. Center Trajectories.

strange attractors. These are significantly more complex and will be briefly discussed later in this chapter.

3.7 Stability: Attracting, Liapunov, and Asymptotic

There are a few types of stability that are often used. The first is known as attracting, where by a fixed point p or a path $p(t)$, is attracting if all trajectories within a given neighborhood of p approach it as time goes to infinity. The second definition of stability is Liapunov Stability, by which a fixed point p or solution path $p(t)$, is Liapunov Stable if all trajectories starting within one neighborhood of p, remain within a different neighborhood of p, for all time. Now, a point or path p is Asymptotically stable if it is both attracting and Liapunov Stable. It would be a good idea to stop here for a moment and ponder the differences.

If p is attracting, this means that all trajectories within a distance δ of p will approach p, as time proceeds. So, given any distance ε, the trajectory will eventually be within ε of p, however, along the way, it may be infiniteley far away. On the other hand, if p is Liapunov Stable, then this means that if a trajectory began within δ of p, then it will be within ε for all time forward. However, I should note that in the case of Liapunov Stability, this means that given an ε there exists a δ such that if a path begins within a distance δ, it remains within ε of the point or path p. However, δ may depend on ε, that is if a smaller ε is given, then a smaller δ may be necessary. This is important because p may be Liapunov Stable, but this does not imply that all paths in a sufficiently small neighborhood of p will approach it in infinite time. They may simply stay a distance of exactly $\varepsilon_1 < \varepsilon$ away for infinite time. For example, paths may form concentric circles around a fixed point p.

A small analogy may sum up the differences. Suppose we have a boy flying a kite. The center (where the boy is standing) would be Liapunov Stable, as long as the string on the kite remained the same length where ε is the distance of the string. This does not mean the kite will ever get closer to the center (the boy), but it does mean that the kite will never get further than the length of the string. It is similar to a fly trapped inside a beach ball, there is no guarantee it will ever reach the center, but surely it cannot leave. On the other hand, suppose the kite started out on an infinitely long string, so it was nowhere near the boy at the center. Then suppose, the string is steadily shortened, the kite will eventually be pulled to the center. However, a large gust of wind may come and it gets really close and then blows away again and the boy lets the string go for a moment and it gets a bit further away, but then continues to be pulled in. In this case, the center is attracting. Now, if the boy had never let the string go, so that the kite continued to get closer and if we made the ε distance equal to the length of the string, then the kite approaches the boy and it never leaves the ε ball, so it is asymptotically stable.

In general, if a point or path is attracting, this means the trajectory will approach it in infinite time. This does not mean that it necessarily gets closer as time moves forward, a trajectory's distance to the fixed point or path need not be monotonically decreasing. This means, a trajectory can get really far away before it comes back. In fact it can come close, then go far away, and then close again, as many times as it wants before it eventually approaches the fixed point or path.

Whereas in the case of Liapunov Stability, the trajectory must stay within some distance of the fixed point or path, the trajectory is trapped by a finite ball around the fixed point or path. However, that is all, it must stay in the ball, but it never has to get any closer to the center than the edge of the ball. In fact, it can simply orbit around the point on the perimeter of the ball and still be Liapunov Stable.

The strongest definition of stability is asymptotic stability which requires that a path approaches the fixed point or stable path and that trajectories with a given distance δ do not get further than ε away for any given $\varepsilon > 0$. This means that the paths have to approach the stable solution and that the paths can not get infinitely far away along the way to approaching the fixed point or stable path. For more mathematically rigorous definitions refer to Strogatz (1994) or Jordan and Smith (1999).

3.8 Linear Stability Analysis

Suppose we have a two dimensional linear system, that is:

$$\dot{x} = f_1(x, y)$$
$$\dot{y} = f_2(x, y)$$

with f_1 and f_2 being linear. Suppose our system is:

$$\dot{x} = ax + by$$

$$\dot{y} = cx + dy$$

then we let

$$\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

so that we now have:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$$

This system is solved as it is solved in most first courses in linear algebra, by first finding the eigenvalues and eigenvectors. In order for $\mathbf{v}$ to be an eigenvector and λ to be an eigenvalue, $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ must be true, so

$$\begin{aligned} \mathbf{A}\mathbf{v} &= \lambda\mathbf{v} \\ &\Rightarrow \mathbf{A}\mathbf{v} - \lambda\mathbf{v} = \mathbf{0} \\ &\Rightarrow (\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0} \\ \mathbf{v} &\in \mathbf{N}(\mathbf{A} - \lambda\mathbf{I}) \end{aligned}$$

where $\mathbf{N}(\mathbf{A} - \lambda\mathbf{I})$ is the null space of $\mathbf{A} - \lambda\mathbf{I}$. However, if $\mathbf{v} \neq \mathbf{0}$ then $\mathbf{A} - \lambda\mathbf{I}$ must be noninvertible and $\mathbf{A} - \lambda\mathbf{I}$ is noninvertible if and only if $\det(\mathbf{A} - \lambda\mathbf{I}) = \mathbf{0}$.

The characteristic equation $f(\lambda)$ is defined as:

$$f(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I})$$

So then λ is an eigenvalue of $\mathbf{A}$ if and only if $f(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \mathbf{0}$.

$$\begin{aligned} f(\lambda) &= \det(\mathbf{A} - \lambda\mathbf{I}) \\ &= \det\begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix} \\ &= (a-\lambda)(d-\lambda) - bc \\ &= ad - a\lambda - d\lambda + \lambda^2 - bc \\ &= \lambda^2 - (a+d)\lambda + (ad - bc) \end{aligned}$$

The trace of a matrix **A** is defined as the sum of its diagonal elements and has the notation, tr(**A**). Substituting in the Trace and Determinant of A, we have the common results:

$$\begin{aligned} f(\lambda) &= \lambda^2 - (a+d)\lambda + (ad - bc) \\ &= \lambda^2 - \text{tr}(\mathbf{A})\lambda + \det(\mathbf{A}) \end{aligned}$$

Now, the characteristic equation is simply a quadratic equation and can be solved using the quadratic rule:

$$\lambda = \frac{\text{tr}(\mathbf{A}) \pm \sqrt{(\text{tr}(\mathbf{A}))^2 - 4\det(\mathbf{A})}}{2}$$

The two eigenvalues are specifically:

$$\lambda_1 = \frac{\text{tr}(\mathbf{A}) + \sqrt{(\text{tr}(\mathbf{A}))^2 - 4\det(\mathbf{A})}}{2}$$
$$\lambda_2 = \frac{tr(\mathbf{A}) - \sqrt{(tr(\mathbf{A}))^2 - 4\det(\mathbf{A})}}{2}$$

The solution to the system is then:

$$x(t) = c_1\mathbf{v}_1 e^{\lambda_1 t} + c_2\mathbf{v}_2 e^{\lambda_2 t}$$

where $\mathbf{v}_1$ is the corresponding eigenvector to λ_1, that is $\mathbf{A}\mathbf{v}_1 = \lambda_1\mathbf{v}_1$ and similarly $\mathbf{v}_2$ for $\lambda_{2.}$ The values of c_1 and c_2 can be found with the initial starting point. At $t = 0$:

$$\begin{aligned} x(t) &= c_1\mathbf{v}_1 e^{\lambda_1 t} + c_2\mathbf{v}_2 e^{\lambda_2 t} \\ &\Rightarrow x(0) = c_1\mathbf{v}_1 e^{\lambda_1 \cdot 0} + c_2\mathbf{v}_2 e^{\lambda_2 \cdot 0} \\ &\Rightarrow x(0) = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 \end{aligned}$$

This system of equations can be solved with some basic row operations from linear algebra. For a review of this material, refer to almost any first year linear algebra text. Now, we can see that as $t \to \infty$ the behavior of $x(t)$ will be dependent upon the values of λ_1 and λ_2, which in turn are dependent on tr(**A**) and det(**A**). Let me define the discriminant D as $D = (\text{tr}(\mathbf{A}))^2 - 4\det(\mathbf{A})$, so that the eigenvalues are:

$$\lambda = \frac{\text{tr}(\mathbf{A}) \pm \sqrt{D}}{2}$$

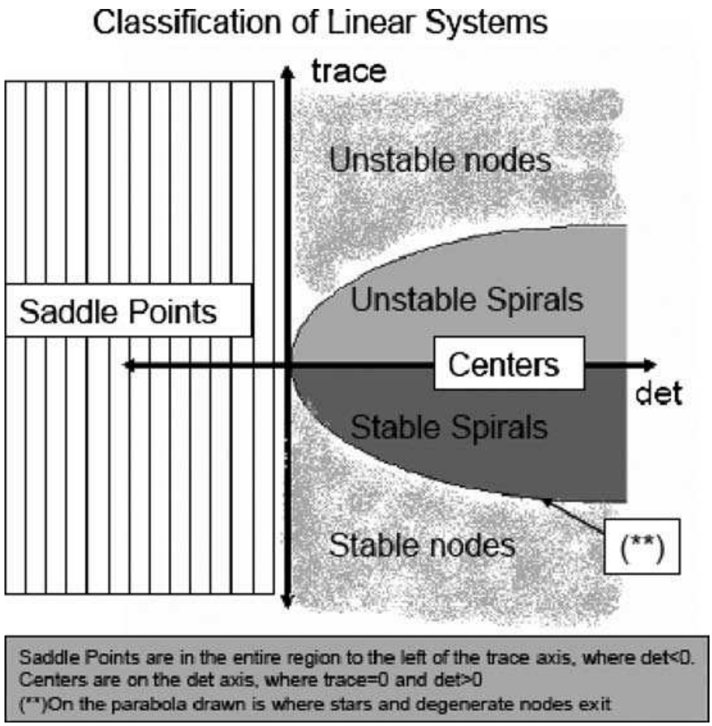


Figure 23. Fixed Point Classification.

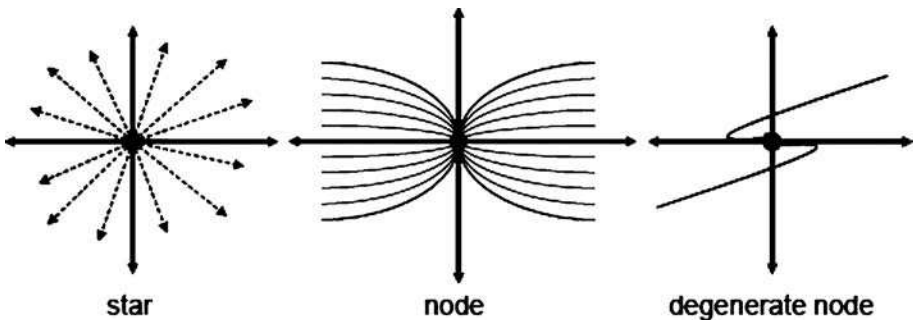


Figure 24. Star, Node, and Degenerate Node.

Figure 23 is often used in order to classify linear systems,where the spirals inside the parabola have $D < 0$ (or $(\mathrm{tr}(\mathbf{A}))^2 - 4\det(\mathbf{A}) < 0$) and the nodes outside the parabola have $D > 0$ (or $(\mathrm{tr}(\mathbf{A}))^2 - 4\det(\mathbf{A}) > 0$). On the parabola itself is where $D = 0$ (or $(\mathrm{tr}(\mathbf{A}))^2 - 4\det(\mathbf{A}) = 0$) and this is where the stars and degenerate nodes exist. In Figure 24, trajectories for a star, node, and degenerate node are given.

4. Linearizing the Nonlinear

Now suppose that we again have a two dimensional system, but now it is nonlinear. That is, we have:

$$\dot{x} = f_1(x, y)$$
$$\dot{y} = f_2(x, y)$$

with f_1 and f_2 now being nonlinear. Suppose, we have a fixed point $a = (x_0, y_0)$ to analyze the stability of a, we will see how a small perturbation from a, effects the trajectory of the system. Our analysis closely follows Strogatz (1994), but this is commonly done using a Taylor series approximation at the fixed point. Thus, using the Taylor series approximation to find $\dot{x}$ at the point perturbed from the fixed point:

$$\begin{aligned}\dot{x}\ (x_0 + \Delta x, y_0 + \Delta y) &= f_1(x_0 + \Delta x, y_0 + \Delta y)\\ &\approx f_1(x_0, y_0) + \frac{\partial f_1}{\partial x}\Delta x + \frac{\partial f_1}{\partial y}\Delta y\\ &\quad + \frac{1}{2!}\left(\frac{\partial^2 f_1}{\partial x^2}(\Delta x)^2 + 2\frac{\partial f_1}{\partial x}\frac{\partial f_1}{\partial y}\Delta x\Delta y + \frac{\partial^2 f_1}{\partial y^2}(\Delta y)^2\right) + \ldots\\ &\approx f_1(x_0, y_0) + \frac{\partial f_1}{\partial x}\Delta x + \frac{\partial f_1}{\partial y}\Delta y \end{aligned} \tag{4}$$

where the final line is an approximation if we assume that since Δx and Δy are small that any terms smaller than them are not significant (these include terms of $(\Delta x)^2$, $(\Delta y)^2$, $\Delta x \Delta y$, and all smaller terms). Similarly, for the second equation:

$$\begin{aligned}\dot{y}\ (x_0 + \Delta x, y_0 + \Delta y) &= f_2(x_0 + \Delta x, y_0 + \Delta y)\\ &\approx f_2(x_0, y_0) + \frac{\partial f_2}{\partial x}\Delta x + \frac{\partial f_2}{\partial y}\Delta y\end{aligned}$$

By noting that $\dot{x}\ (x_0, y_0) = f_1(x_0, y_0)$, the above equations can be manipulated as:

$$\begin{aligned}\dot{x}\ (x_0 + \Delta x, y_0 + \Delta y) - \dot{x}\ (x_0, y_0) &= f_1(x_0 + \Delta x, y_0 + \Delta y) - f_1(x_0, y_0)\\ &\approx \frac{\partial f_1}{\partial x}\Delta x + \frac{\partial f_1}{\partial y}\Delta y\end{aligned} \tag{5}$$

Similarly,

$$\dot{y}(x_0 + \Delta x, y_0 + \Delta y) - \dot{y}(x_0, y_0) = f_2(x_0 + \Delta x, y_0 + \Delta y) - f_2(x_0, y_0)$$
$$\approx \frac{\partial f_2}{\partial x}\Delta x + \frac{\partial f_2}{\partial y}\Delta y \tag{6}$$

From 5 and 6, comes the following linearized system of equations:

$$\dot{x}(x_0 + \Delta x, y_0 + \Delta y) - \dot{x}(x_0, y_0) = \dot{\Delta x} = \frac{d}{dt}(\Delta x) = \frac{\partial f_1}{\partial x}\Delta x + \frac{\partial f_1}{\partial y}\Delta y$$
$$\dot{y}(x_0 + \Delta x, y_0 + \Delta y) - \dot{y}(x_0, y_0) = \dot{\Delta y} = \frac{d}{dt}(\Delta y) = \frac{\partial f_2}{\partial x}\Delta x + \frac{\partial f_2}{\partial y}\Delta y$$
$$\Rightarrow \begin{pmatrix} \dot{\Delta x} \\ \dot{\Delta y} \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$$

Now we have a linear approximation of the system and we can do the same analysis as before, but with some dangers to beware of! Since we removed the higher order terms from the taylor expansion, sometimes the linearization is not correct. The cases that are "borderline" in the classification diagram above are those where the linearization cannot be trusted. So, the linearization is good for nodes, spirals, and saddles but not centers, stars, or degenerate nodes.

This should make sense because the borderline cases are delicate. A center is like a spiral that lines up just perfectly and a star is like a node that is perfectly straight. However, if the spiral needs to line up perfectly or the node must be perfectly straight, then the small difference of the higher order terms may destroy these rare cases. A rigorous explanation of the failure of the linearization is quite complex and beyond the scope of this introductory chapter, but refer to Strogatz (1994) for a more detailed explanation.

5. Strange Attractors

Most systems give rise to an attractor that is a fixed point or loop. However, some attractors are far more complex and are known as strange attractors. One of the most popular and also the first strange attractor to be discovered is the Lorenz Attractor shown in Figure 25. Strange attractors can occur in continuous or in discrete time dynamical systems. However, it should be noted that in the continuous time case, they can only occur if the dimensionality of the system is

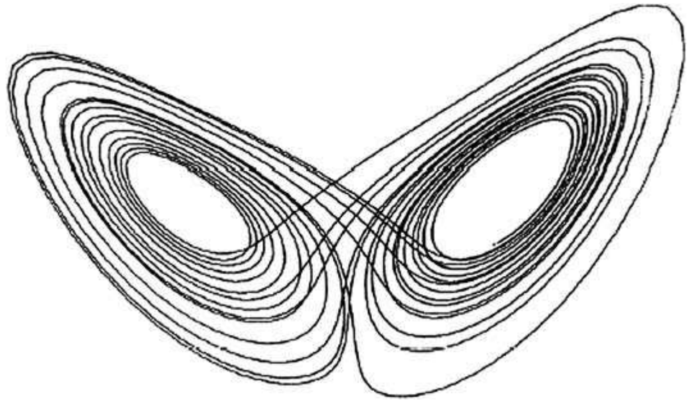


Figure 25. Strange Attractor (Stewart, 1989).

three or greater, whereas in the discrete case, they can occur in even the single dimension case (according to the Poincare-Bendixson theorem explained later).

6. Trajectories

It should be noted that a trajectory is simply the path taken in the phase diagram, given an initial starting position. So if a system begins in a different position, then it will have a different trajectory. Precisely those systems that are interesting are those that are chaotic. These systems have trajectories that are very sensitive to initial conditions. That is if the system starts at x, by some time t, it will be radically different than if it had started at y, even if x and y are very close.

When we say "very close" or "radically different," this often means numerically close or different, but for different systems, "close" may take on different definitions. For example, Lorenz coined the term, "Butterfly Effect."The basic idea was that if a butterfly flaps its wings then in some far off time, it may cause a tornado in Texas. In this example, the state of the world without the butterfly flapping its wings is "close" to the state where it does. Similarly, the state without the Tornado is far from the state with it simply because in this example we perceive the butterfly flapping its wings to be a small event and the Tornado to be a large one. A different metric may be used on different state spaces.

In a more practical sense, we may measure several variables such as temperature, humidity, and wind speed in order to find the current state of the weather. From, this initial state we may predict it will rain five days from now. But, instead it snows because our measurements today were slightly off.

So, two states that are very similar today may lead to states that are not only numerically different but are qualitatively different as well.

Since different trajectories could lead to different states, it may be difficult to understand the behavior of some systems by simply looking at one or two trajectories. Some physical systems may be modeled by equations that are not known perfectly, that is the parameters are estimates or the equations are not exact. In this case, if the trajectories are not close enough for a set of initial conditions, then it is difficult to describe the behavior because these model approximations may be enough to cause radically different trajectories. For such systems, it may be difficult to predict the topology of trajectories or their stability.

Some systems may have different sets of trajectories exhibiting different properties. One set could lead to a fixed point and another to a loop, for example. In a more complex system, you may be missing a lot if you simply look at one trajectory. Some systems are very complex and make it difficult to describe or even know all of the classes of trajectories.

Other systems may have classes of trajectories that radically change as the parameters change. So, looking at one trajectory and one parameter value will just be one case. This gives rise to the investigation of bifurcations, where a change in a parameter may cause a significant change in the behavior of the system. As a system's dimensionality increases, the complexity of its behavior can get harder to describe because there can exist more and more classes of trajectories which may act differently for different parameter values. Since we can only plot a system in two dimensions on paper and three on a computer, it gets very hard to imagine fully the behavior of higher dimension systems.

7. Bifurcations

Sometimes the trajectory or solution of a dynamical system is dependent on a parameter. As the parameter changes, the trajectories will surely change, however, the behavior or classification of the trajectory could change as well. For example, if a trajectory depends on a parameter α and when $\alpha < \alpha_o$ the structure of the phase space is a fixed point, but then when $\alpha \geqslant \alpha_o$, the phase space changes its structure to a limit cycle, then the system has experienced a bifurcation at α_o. Bifurcations can occur by creating or destroying an attractor or repeller, or by changing an attractor or repeller from one type to another (e.g. fixed point to limit cycle, or attractor to repeller).

7.1 Saddle-Node Bifurcation

The simplest example of a bifurcation is the saddle-node bifurcation where equilibrium points are either created or destroyed. A nice example taken from

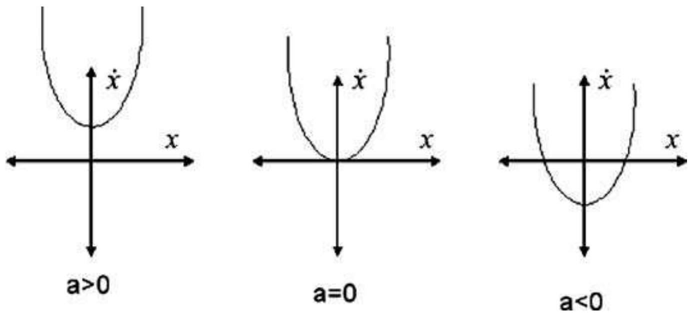


Figure 26. Saddle Node Bifurcation.

Strogatz (1994) is:

$$\dot{x} = a + x^2$$

Remember that a fixed point or equilibrium point occurs where there is no flow, that is where $\dot{x} = 0$. This means the equilibrium points may be found by finding the roots of the equation $0 = a + x^2$. However, the number of roots depends on the value of a:

$a < 0$	There are two real roots, $x = \sqrt{-a}$ and $x = -\sqrt{-a}$
$a = 0$	There is one real root, $x = 0$
$a > 0$	There are no real roots

This means that if we let $a > 0$, then there are no equilibrium points, but if we decrease a, then an equilibrium point is created when $a = 0$. Then, as a is decreased further, the equilibrium point immediately splits into two points, when $a < 0$. Thus there is a bifurcation at $a = 0$ because two equilibrium points are created or destroyed, depending on which way you look at.

Since a fixed point exists when $\dot{x} = 0$, we can see that the roots in Figure 26 are the fixed points. None exist for $a > 0$, one at $a = 0$, and two for $a < 0$. Refer to Figure 26 to see the graph of a saddle node bifurcation.

7.2 Transcritical Bifurcation

A transcritical bifurcation is one in which the equilibrium point changes is stability from stable to unstable or vise versa as a parameter varies. A good example of a transcritical bifurcation again from Strogatz (1994):

$$\dot{x} = ax - x^2$$

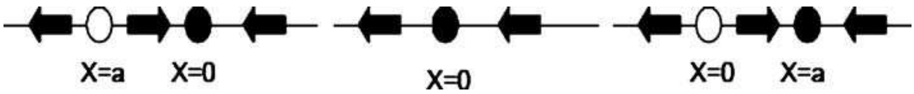


Figure 27. Transcritical Bifurcation: Stability Swap.

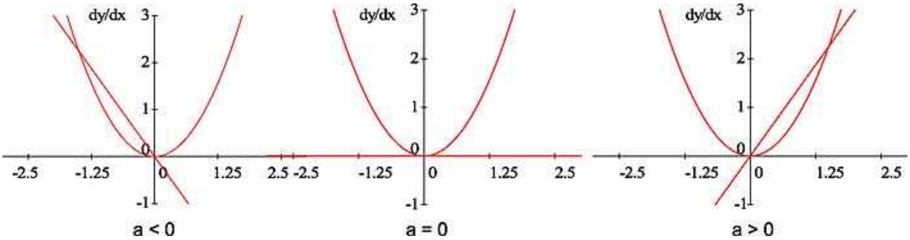


Figure 28. Transcritical Bifurcation.

Now, when:

$a < 0$	$x = a$ is an unstable fixed point and $x = 0$ is a stable fixed point
$a = 0$	$x = 0$ is a single half stable fixed point because the two fixed points merged
$a > 0$	$x = a$ is a stable fixed point and $x = 0$ is an unstable fixed point

Thus as a increased, $x = 0$ was a fixed point that lost its stability. The fixed point $x = a$ moved to the right as a increased and when it crossed $x = 0$, they merged briefly as a half stable fixed point and then when it passed $x = 0$, it took $x = 0$'s stability with it and left $x = 0$ as an unstable fixed point. Hopefully, Figure 27 will help clarity this point.

In Figure 27, again we can see there are two roots for $a < 0$ as well as $a > 0$, but only one for $a = 0$. Now, in order to analyze the stability we must look a bit deeper at this diagram. Since $\dot{x} = ax - x^2$, when the line is above the parabola $ax > x^2$ and $\dot{x} > 0$, so the flow is to the right, yet when the line is below parabola, the flow is to the left.

From the Figure 28, we can see that when $a < 0$, the flow is away from $x = a$, making this the unstable point, where as the flow is toward $x = 0$, making it the stable point. When $a = 0$, the flow is to the left on both sides, here $x = 0$ is called semi-stable. Then when $a > 0$, the fixed points swap stabilities making $x = 0$ unstable and $x = a$ stable.

There are many other bifurcations studied in dynamics. Virtually any parameterized system could have a bifurcation and many can be unique in there own way. We gave a single example of a saddle-node bifurcation and a transcritical bifurcation. These are simply classes of bifurcations, but there are many types which may not even be classified. In any event, two other types of bifurcations not discussed were the Pitchfork Bifurcation and Hopf Bifurcation. Refer to Strogatz (1994) for a more in depth explanation of more bifurcations.

8. Poincare-Bendixson Theorem

The broad implications of this theorem are that all trajectories in a two dimensional continuous time system must converge to either a fixed point or a limit cycle. So chaotic behavior (strange attractors) can only occur in systems with more than two dimensions or discrete time systems.

More specifically, the theorem says that if there exists a closed and bounded region R, often called a trapping region, such that R contains no fixed points, then all trajectories inside R must either be a closed orbit or spiral toward one.

9. Logistic Map

This is a population model first used by Pierre Francois Verhulst. The biologist Robert May was the first one to shown the applications to Dynamical Systems and Chaos Theory in his 1976. The logistic map is very simple in that there is only one equation in the dynamics. Yet this single nonlinear dynamic equation can give rise to very chaotic behavior.

The logistic map can be written as:

$$x_{t+1} = r x_t (1 - x_t)$$

where:

$$x_t = \text{population in year } t \text{ (between 0 and 1)}$$

$$x_0 = \text{population in year 0 (between 0 and 1)}$$

$$r = \text{rate of starvation and reproduction (positive number)}$$

The fate of the population is quite dependent on r:

$0 \leqslant r \leqslant 1$	Population will die regardless of initial population ($x_t \to 0 \forall x_0$)
$1 < r \leqslant 2$	Population will quickly converge to $\frac{r-1}{r}$
$2 < r \leqslant 3$	Population will converge to $\frac{r-1}{r}$, but at a slower rate
$3 \leqslant r \leqslant 1+\sqrt{6}$	Population will oscillate between two values (dependent on r but not x)

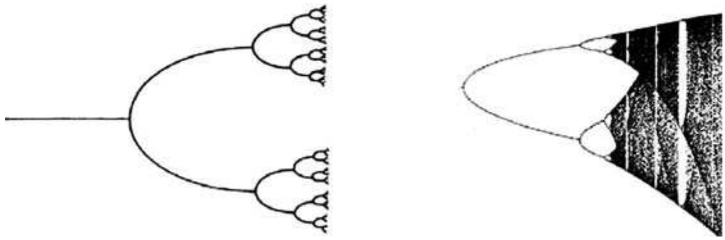


Figure 29. The Logistic Map (Stewart, 1989).

As the ranges increase the number of values x_t oscillates between, keeps doubling. It starts with two values and goes to 4 then 8 and 16, and so on. Until about $r \geqslant 3.57$ where the logistic map become chaotic. In this region, a slight difference in the initial condition can cause drastically different trajectories. Then at $r \geqslant 4$, x_t will diverge for almost all initial values.[4]

The logistic map is thus sensitive to the initial conditions for some values of r and not sensitive for others. The logistic map is an example of a dynamical system that goes through bifurcations multiple times. One of the amazing things to note is that such an interesting example appears as a one dimensional problem. This is because it is an iterative map (as opposed to a system of differential equations), which can display much more complex behavior with less dimensions.

10. Routes To Chaos

Most systems for some parameter values have nonchaotic behavior, that is limit cycles, closed orbits, and fixed points. Then as a parameter is changed, the system becomes chaotic. The way in which it becomes chaotic is referred to as a "route to chaos." There are various routes to chaos that a system may take. In fact, for different parameter values, the system make take different routes. The important thing to realize is that systems become chaotic in similar ways, that is, there are classes called "routes to chaos." The study of chaos theory is still young and quite open to research, so there still may be many routes yet to be discovered but nonetheless systems can be grouped by how they become chaotic. For further reading on the routes to chaos, refer to Hilborn (1994).

Hilborn (1994) describes *period doubling* as a route to chaos, in which a system has a limit cycle that becomes unstable as a parameter changes. Then, at some point the period doubles, then it doubles again, and keeps doubling

[4] Figure 29 taken from Stewart (1989) should bring clarity to this chaotic system.

until the period becomes infinite so that the trajectory never makes a second trip. At this point the system is chaotic.

Hilborn (1994) describes *intermittency* as a route to chaos when a system has periodic behavior with irregular chaotic (non periodic) bursts of behavior. Then, as a parameter is changed, the bursts of irregular behavior become longer and more frequent, until the behavior is completely irregular at which time it is said to be chaotic.

These are just two ways in which a system may become chaotic. For a more rigorous description of these routes to chaos as well as others, refer to Hilborn (1994) or any other text in nonlinear dynamics.

11. Concluding Remarks

After reading this chapter, the reader should certainly be convinced that networks are all around us. The topology of these networks can certainly effect the network's vulnerability as well as its capacity to transport commodities. While investigating any objective over a network, it is important to consider the network's topology.

The study of a network's topology and its effects is a study in its own. However, topology can also effect the dynamics of systems. Dynamical Systems and the theory of chaos has certainly gained much attention in recent years. Now, there may be some systems governed by dynamics such as epidemics that will certainly be effected by the topology of the network that it is acting on.

References

A. L. Barabási. Linked: The New Science of Networks, Perseus, Cambridge, MA, 2002.

R. Albert, H. Jeong, and A.-L. Barabási. Attack and error tolerance of complex networks. Nature, 406 (2000), pp. 378–382.

J. Balthrop, S. Forrest, M. Newman, and M. Williamson. Technological networks and the spread of computer viruses. Science, 304, (2004) pp. 527–529.

M. Beckman, C. B. McGuire, and C. B. Winsten. Studies in the Economics of Transportation. Yale University Press, 1956.

N. L Biggs, E. K. Lloyd, and R. J. Wilson. *Graph Theory: 1736–1936*. Oxford, England. Clanderon Press. 1976.

S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D. U. Hwang. Complex networks: Structure and dynamics. Phys Rep 424 (2006) pp. 175–308.

P. Erdős and A. Rényi. *On random graphs*, Publicationes Mathematicae Debrecen, 6 (1959), pp. 290–297.

P. Erdős and A. Rényi. *On the evolution of random graphs*, Magyar Tud. Akad. Mat. Kutató Int. Közl., 5 (1960), pp. 17–61.

P. Erdős and A. Rényi. *On the strength of connectedness of a random graph*, Acta Math. Acad. Sci. Hungar., 12 (1961), pp. 261–267.

S. Friedberg, A. Insel, L. Spence. Linear Algebra, Prentice Hall, Englewood Cliffs, NJ. 2003.
J. Guare. Six Degrees of Separation: A Play, Vintage, New York, 1990.
R. C. Hilborn. Chaos and Nonlinear Dynamics Oxford University Press, New York, NY. 1994.
D. W. Jordan and P. Smith. Nonlinear Ordinary Differential Equations, Oxford University Press, New York, NY, 1999.
F. Karinthy. Chains, in Everything is Different, Budapest, 1929.
S. Milgram. The small world problem, Psych. Today, 2 (1967), pp. 60–67.
M. E. J. Newman. The Structure and Function of Complex Networks, Society for Industrial and Applied Mathematics, 45 (2003), pp. 167–256.
M. E. J. Newman. Spread of epidemic disease on networks, Phys. Rev. E, 66 (2002).
P. Nijkamp and A. Reggiani. *Interaction, evolution and chaos in space*. Springer-Verlag, Berlin, 1992.
A. C. Pigou. *The economics of welfare*. Macmillan, 1920.
G. Perakis. The price of anarchy when costs are non-separable and asymmetric. In Proceedings of the 10th Conference on Integer Programming and Combinatorial Optimization (IPCO), volume 3064 of Lecture Notes in Computer Science, pages 46–58, 2004.
T. Roughgarden. É. Tardos, How bad is selfish routing? J. ACM 49 (2) (2002) pp. 236–259.
T. Roughgarden (2002), Selfish Routing, PhD dissertation, Cornell University.
I. Stewart. Does God Play Dice?, Basil Blackwell Ltd, Malden, MA, 1989.
S. H. Strogatz. Nonlinear Dynamics and Chaos, Addison–Wesley, Reading, MA, 1994.
H. P. Thadakamalla, U. N. Raghavan, S. Kumara, and R. Albert. Survivability of multiagent-based supply networks: A topological perspective. IEEE Intell. Syst., (2004), 19(5), 24–31.
D. J. Watts and S. H. Strogatz. Collective dynamics of “small-world” networks, Nature, 393 (1998), pp. 440–442.
D. J. Watts. Proc. Natl. Acad. Sci. USA 99 (2002) 5766.
Bureua of Transportation Statistics: National Transportation Statistics. Online. October 2006 http://www.bts.gov/publications/national_transportation_statistics/

Chapter 3

Modeling Large Scale and Complex Infrastructure Systems as Computable Games

Terry L. Friesz,[1] Reetabrata Mookherjee,[2] and Srinivas Peeta[3]

[1] *Harold and Inge Marcus Chaired Professor of Industrial Engineering, The Pennsylvania State University, University Park, PA 16802, USA, E-mail: tfriesz@psu.edu*

[2] *Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802, USA, E-mail: reeto@psu.edu*

[3] *School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA, E-mail: peeta@purdue.edu*

Abstract Infrastructure systems for generalized transportation – such as goods, passengers and water – take the form of networks. These networks typically have interdependencies which are not addressed in engineering practice. In order to make efficient policy regarding an infrastructure system, the impacts of that policy on other interdependent infrastructure systems must be understood. The combination of the different layers of the interconnected infrastructure network may be thought of as a system of systems representing the grand infrastructure system. Users of the system of systems may be thought of as agents competing for the limited capacities of the network layers. Dynamic game theory is a natural method for modeling systems of systems in an effort to make better infrastructure decisions. However, to be of use, these models must be computable and thus some different solution techniques for general equilibrium models are discussed.

Keywords: infrastructure; system of systems; spatial equilibria; dynamic games

1. Introduction

Although the main goal of this book is to investigate and conjecture about the relevance of network science and nonlinear science to the study and management of infrastructure, the emerging notion of a *system of systems* is an allied paradigm that may be of great assistance in achieving that goal. Despite the fact that abstract and formal definitions may be given of both a system and a system of systems, we believe a more helpful introduction to those notions may be obtained from a rather informal look at infrastructure networks and the reasons they may be considered a system of systems. In particular, we argue below that individual traditional as well as cyber infrastructure systems naturally take the form networks that are coupled to one another in a variety of ways. Such coupled networks are *per force* a system of systems.

Similarly, the theory of dynamic games provides a natural language for articulation of the system of systems view of infrastructure. In particular, most difficult infrastructure decisions, of both a tactical and strategic nature, occur because of prior, current or anticipated changes in individual infrastructure systems and the competition that occurs among users of the bounded infrastructure services those systems provide. Dynamic game theory is a fundamental paradigm for modeling such dynamic competition.

2. Infrastructure Services as Generalized Transportation

We follow Peeta et al. (2005) in acknowledging a growing awareness that all infrastructure technologies involved in generalized transportation – broadly defined to include goods, passenger, message, data, water and energy flows – are organized as networks and that these networks are interdependent and can be thought of as coupled layers of a grand infrastructure network. In addition, as Peeta et al. (2005) point out, recent advances in information technology constitute an enabling information infrastructure that provides synergism to inter-layer couplings among traditional individual infrastructure networks. Yet, the interdependence of infrastructure networks is generally ignored in engineering practice. That is to say, infrastructure network (IN) planning and design is done one network at a time with only the most cursory acknowledgment of the couplings that exist among INs. This inconsistency, once observed, begs the question: What is the nature of the interdependencies among INs and can those interdependencies be included in engineering analyses?

In that each IN is a type of system, all the individual infrastructure networks together with their linkages to one another constitute a system of systems. Seen in this way a system of systems is a natural perspective to take when studying infrastructure. Moreover, this observation is not really new, for such systems have been discussed in the transportation modeling literature using alternative names. In particular, Sheffi (1985) refers to coupled, multi-modal transport networks as *hypernetworks*, and it is just a short step from his perspective to the more general notion of an *infrastructure system of systems* (ISS). More recently Nagurney (2006) has used the name *supernetwork* to describe coupled networks of energy, transport and other physical flows that echo the system of systems perspective of Peeta et al. (2005).

It is our contention in this paper that in order to make coherent policy regarding the critical infrastructures of our national economy it is necessary to answer the following question: How can interdependencies among individual INs be expressed mathematically so that richer and more informative models to support IN planning and design may be formulated and numerically solved? A companion question, also needing an answer is: Will models that capture the system of systems (SOS) perspective that seems so natural for description of infrastructure systems be computable?

3. Origin of Regional and National Infrastructure Systems of Systems

There are five main sources of interdependency or *coupling* among generalized transportation networks that comprise infrastructure systems. These are:

1. **Physical Interdependencies**. Infrastructure networks are sometimes coupled by virtue of shared physical flow rights of way leading to joint capacity constraints. Telecommunications and data networks are one example. Passenger and goods networks are another.
2. **Budgetary Interdependencies**. Most infrastructures associated with generalized transportation networks involve some degree of public financing so that the financing of one IN either directly or indirectly affects the financing of others.
3. **Market Interdependencies and Spatial Economic Competition**. With the increasing globalization of the world's economy and the trend toward ever more intelligent infrastructure, spatially separated supplies and demands for the services and goods exchanged over INs generally form a single global competitive market and, thereby, influence one another even when other explicit interdependencies are not manifest. Moreover, because of the public good aspect of many INs, numerous governmental regulations exist and are emerging that control both intra- and inter-layer aspects of the spatially extended economic competition that occurs via INs.
4. **Information Interdependencies**. With recent advances in enabling information technology, comprehensive data *and* information infrastructures are commonly available. As a consequence, database sharing and information exchange among individual INs provides synergism and cost-efficiency. For example, solution strategies to improve freight movements and the quality of passenger travel share common databases on traffic conditions across vehicular transportation networks. Similarly, urban water and energy

utilities may share information on the socioeconomic characteristics of individual households to more consistently predict future demands.

5. **Environmental and Congestion Externalities**. It is clear that transportation networks manifest congestion. It is also clear that transportation systems as well as power generation and distribution systems may generate undesirable air and water effluents. These externalities are not confined to the infrastructure network that created them; rather, they directly impact other individual infrastructure networks. A simple example is that effluents from power generation may contaminate water supplies, thereby increasing the costs of delivering potable water to citizens.

When these sources of coupling are recognized, a natural conceptual model of INs begins to emerge. Specifically, INs may be viewed as multilayer networks with coupling constraints among the layers; these layers may be arranged in various hierarchies, depending on their societal and engineering function. The resultant multi-layer, coupled infrastructure networks constitute a system of systems.

4. Defining A System of Systems

Recent literature offers several definitions of a system of systems (SoS). Notable among these are Crossley (2004), Keating et al. (2003), Levis (2004) and Sage and Cuppan (2001). A system of system (SoS) is thought of by all these authors as a collection of systems, each of which is capable of independent operation, that must interact to achieve their purposes or gain value none can fully realize alone. Like a single system, an SoS is a collection of components interacting to fulfill one or more functions. But the constituent systems of an SoS can perform useful functions alone – something components of a single system cannot – and removal of any system from an SoS need not prevent its continued operation. So in terms of our infrastructure terminology the systems of an SoS are INs and the SoS itself is the coupled collection of INs.

A SoS is not a rare or seldom seen event. Rather, many societal needs are met by SoSs. In each case, component entities interact with each other, but they have operational and managerial independence, and their geographic extent limits their interactions to flows of goods, customers, information, and/or money. They usually operate in an environment of considerable risk and uncertainty, and may exhibit unpredicted emergent behaviors as consequences of their interactions. Some SoSs are mainly technological, e.g. the FAA's Traffic Alert and Collision Avoidance System (TCAS) in which hardware and software on independently operating planes exchange information and locations to produce alerts with little human intervention. Others are primarily human enterprise SoSs where human decision-making and interactions have a central role; e.g., the congestion of an urban road network is governed almost solely by the decisions of individual drivers. Most SoSs fall along a spectrum between both extremes. Similarly, constituents within an SoS may share a common goal or mission, each constituent may be operating entirely in its own interest, or a combination of cooperating and competing systems may exist.

5. Foundation Disciplines for a Science of Infrastructure

As has been said above, the complete infrastructure of a city, region, or country may be viewed as a set of generalized transportation networks coupled to one another in the various ways discussed in Section 3. The key generalized transportation networks, each itself a system of systems and which taken together form a grand infrastructure SoS, are readily identified: transportation, telecommunications, water distribution, energy distribution, supply chains and cyberinfrastructure. Although extensive statistics could be presented to establish the strategic importance of these generalized transport networks, there is no real doubt concerning their vast scope and influence on any city, region, or country.

Instead we want to stress that there are two overlapping perspectives regarding SoS infrastructure scholarly inquiry: (1) that system architecture and qualitative process engineering should be the primary foci of SoS research and development, and (2) that extension of mathematical tools from operations research, engineering design, economics, nonlinear science and network science are needed to deal with the challenges presented by SoSs. This distinction is fundamental. In this paper, and for that matter in this book, we take the second point of view: namely, that of adapting mathematical tools drawn from engineering design, operations research, economics, nonlinear science and network science.

To clarify our point of view, it is helpful to offer brief definitions of the foundation disciplines cited above. *Engineering Design* concentrates upon the product development process for engineered systems; this encompasses techniques to generate and track requirements, generate and improve design concepts through trade studies and parametric optimization based on predictive models. *Operations Research* is the collection of theoretical results and algorithms for describing and managing engineered systems, usually with an emphasis on efficiency. As such operations research is very intimately connected to *Game Theory*, which is concerned with mathematical descriptions of various types of competition. *Nonlinear Science* is the modern version of dynamic systems; it focuses on unusual behaviors and gives special attention to notions of stability, complexity and phase transition. *Network Science* is the newest of the foundation disciplines; it is leading the development of a general theory of networks that has uncovered immense similarities in diverse settings. *Computer Science* addresses issues including computational architectures, very large scale distributed computing, data extraction and mining, standards for data exchange, and secure multi-party computation.

5.1 Nonlinear Science

The special, complex behaviors of key state variables over time that is emphasized in nonlinear science include bifurcation, catastrophe, chaos and strange attractors. The appearance of such unusual behaviors in an engineered system is usually viewed as highly problematic and undesirable. However, seen from another angle, nonlinear science can tell us where and how to look for problematic behaviors in SoSs. Once these problematic features are found, intervention schemes that are technologically feasible and functionally desirable for the application being studied may be designed and implemented by engineers.

5.2 Network Science

In many ways the story is the same for network science, which identifies and describes certain recurrent self-organizing behaviors in networks. Some of this organizing behavior leads to networks that are inequitable in that the "winner" of a competition conducted on a network may take all the rewards and leave fellow competitors with none. Alternatively, sometimes network structures evolve that are extremely secure but highly cumbersome to use, while in other cases network structures evolve that are very user friendly but highly vulnerable. Network science tells us where and when such network configurations will emerge. In so doing network science is evidently crucial to designing intervention schemes appropriate to a network controller's stated objectives.

Unquestionably, the most striking discovery at the heart of network science is that many seemingly diverse networks have very similar topological characteristics that arise from the aggregate behavior of a large number of users whose individual behaviors range from wholly non-cooperative to fully collusive. Among the similarities created by the behaviors of large ensembles of network agents and shared by nearly all complex networks are: (1) weak coupling among certain subnetworks that are internally nearly fully connected; (2) occurrence of dominant hubs with disproportionate numbers of connected links; (3) appearance of disparate fitness levels of individual nodes leading to winner-take-all outcomes; (4) power law distributions for which small events co-exist with large events; (5) phase transitions from disorder to order marked by power law distributions; (6) self-similarity and lack of a natural scale;

(7) strong correlation between resilience and self organization; (8) strong correlation between topological robustness and exposure to attack; and (9) emerging self-awareness and organic nature of complex networks. Most of the features of networks from this list have been identified through the adaptation of mathematics and measurement techniques from statistical mechanics and the physics of very large systems; see Albert and Barabasi (2002), Gupta and Campanha (2003), Newman (2003), Parisi (2004) and Venkatasubramanian et al. (2004).

6. Cybernetworks and Cyberinfrastructure

The U.S. National Science Foundation (NSF) has coined the name *Cyberinfrastructure* (CI) to describe the extension, made possible by high-performance computing and advanced information technologies, of the traditional infrastructure associated with scholarship. Exploiting the full potential of CI is expected to bring about significant increases in national productivity through enhanced educational and research efficiency. Accordingly, for a fully modern and general investigation of infrastructure as a system of systems, one must make certain that one of the subsystems studied is cyberinfrastructure. This is a challenging task in that somehow the general, features education and training, distance collaboration and data/model sharing among the infrastructure scholars must be given a mathematical representation.

The network science literature already contains a number of contributions on modeling the type of dynamic information networks and social networks that comprise cyberinfrastructure. For example, much has been written about the affiliation, co-authorship and citation-based networks of scholars, including Lotka (1926), Kretchschmer (1994), Persson and Beckmann (1995), Watts and Strogatz (1998), Albert et al. (1999), Newman (2003), and Watts (2003). These works, however, are concerned primarily with the identification and description of social and information networks that arise in a given environment; they do not provide a direct means of designing cyberinfrastructure networks.

7. Computable Games and Multi-Layer Infrastructure Networks

The performance and evolution of SoSs can be significantly influenced by the decisions taken by individuals or groups at various levels in the subsystems. Typically the goals or intent of the various levels are in conflict, and the situation being modeled is referred to as “competitive” or “non-cooperative”. For example, non-cooperative decision-making strategies compounded by the advances in information technology, can manifest as cascading system failures, as has been illustrated by the massive power blackout in 2003 in the northeastern United States where a specific energy delivery system became the critical weak link in the power grid SoS. The literature on modeling competitive systems has converged on the paradigm of non-cooperative game theory as the fundamental modeling perspective. Non-cooperative games, in addition to being apt descriptions of competition among agents, enjoys a rich and compelling literature on efficient computation.

The computable game theory literature has focussed mainly on determining static equilibria. However, much of the insight gained in the last 20 years concerning computation of game-theoretic equilibria may be extended to dynamic settings in which intelligent infrastructure systems operate. Peeta et al. (2005) have researched multilevel network games that correspond to systems of systems and in which decision-makers associated with distinct tiers (systems) – such as manufacturing, distribution, retailing, and the like – may compete within own tiers (systems) yet cooperate with some if not all non-own tiers (systems). Examples of such SoS games include supply chains, financial networks, and electric power distribution.

We discuss below how both equilibrium and the dynamic multi-layer IN models can be coupled to traditional engineering notions of capital budgeting and network design to create a new generation of IN mathematical and software tools. These mathematical and software tools are intended to provide a foundation for new advanced engineering decision support systems for IN planning and design.

7.1 Background on Spatial Computable General Equilibrium Models

Spatial computable general equilibrium (SCGE) models are spatial extensions of computable general equilibrium (CGE) models [see, for example, Scarf and Hansen (1973) and Mathiesen (1985a and 1985b)]. The spatial extension is achieved by explicit mathematical representation of generalized transportation networks over which goods and services move. As we have already mentioned, the transportation networks of interest in this paper are *generalized transportation networks* (GTNs) whose flows correspond to goods, passenger, message, data, water and energy movements.

We next construct a generalization of the CGE model to describe an economy comprised of spatially separated markets that are interconnected by a generalized transportation network over which goods, passengers, messages, data, water and energy flow; these coupled markets form a perfectly competitive economy. We refer to each of these types of flow as a *commodity flow* (or simply *commodity* for short). Moreover, each commodity $i \in \mathbf{N}$

corresponds to a specific infrastructure technology and is confined to a specific infrastructure network whose graph we give the designation IN_i, so that the grand (multicommodity) infrastructure network IN obeys

$$IN = \bigcup_{i=1}^{|\mathbf{N}|} IN_i$$

In the model developed in this section some simplifying assumptions are made to streamline the exposition. The most important of these are that all generalized transportation flows experience congestion and that routes are determined from a Cournot-Nash non-cooperative game theoretic equilibrium. In the general case, we would allow different behavioral principles to describe routing on each subnetwork IN_i; more discussion of this issue of multiple network behavioral principles is provided below.

We will need the following notation:

$\mathbf{R}$	the set of locations indexed by r and s
$\mathbf{M}$	the activity set for every location indexed by j
$\mathbf{N}$	the commodity set for every location indexed by i
$\mathbf{L}$	the set of network links indexed by a
Λ	the O-D pair/path incidence matrix
h_p^i	the flow of commodity i on path $p \in \mathbf{P}$
f_a^i	$\sum_{p \in \mathbf{P}} \delta_{ap} h_p^i$, the flow of commodity i on link $a \in \mathbf{L}$
$\tau_a^i(\mathbf{f})$	the unit cost of transporting commodity i over link $a \in \mathbf{L}$
$c_p^i(\mathbf{h})$	$\sum_{a \in \mathbf{L}} \delta_{ap} \tau_a^i(\mathbf{f})$, the unit cost of transporting commodity i over path p
u_{rs}^i	$\min_{p \in \mathbf{P}_{rs}} c_p^i(\mathbf{h})$, the minimum cost of transporting commodity i between O-D pair (r, s)
$\mathbf{P}_{rs}$	the set of paths for O-D pair (r, s)
Δ	the link-path incidence matrix
δ_{ap}	an element of Δ, where $\delta_{ap} = 1$ if link a belongs to path p and $\delta_{ap} = 0$ otherwise
λ_p^{rs}	an element of Λ where $\lambda_p^{rs} = 1$ if path p connects O-D pair (r, s) and $\lambda_p^{rs} = 0$ otherwise
y_j^r	the output level of activity j in location r
π_i^r	the supply price of commodity i in location r
b_i^r	the initial endowment of commodity i in location r
$d_i^r(\pi)$	the demand function in location r for commodity i
$a_{ij}^{rs}(\pi, \mathbf{u})$	the input-output coefficient of activity j in location s relative to commodity i produced in location r

$\mathbf{A}(\pi, \mathbf{u})$ the price and transport cost dependent activity analysis matrix
T_i^{rs} the demand for transportation of commodity i from origin r to destination s

In general, to refer to a vector we simply drop the superscripts and/or subscripts. For example, h_p^i denotes the flow of commodity i on path $p \in \mathbf{P}$ and $\mathbf{h} \in \Re_+^{|\mathbf{P}| \times |\mathbf{N}|}$ denotes the commodity path flow vector. It is also important to observe that, in the notation above, we have considered a general case allowing for spatially differentiated commodities. Spatially homogeneous commodities may be represented by a suitable choice of coefficients in the $A(\pi, \mathbf{u})$ matrix. In order to simplify notation, final demand transactions are assumed to occur locally.

The introduction of flow dependent generalized transportation costs, $\mathbf{u}(\mathbf{f})$, implies that there is a difference between supply and demand prices in the economy. This difference in prices of commodities in different locations induces generalized transportation demands. Clearly, the generalized transportation demand is elastic in the sense that $\mathbf{A}(\pi, \mathbf{u})$ is dependent on $\mathbf{u}$. Furthermore, dependency of $a_{ij}^{rs}(\pi, \mathbf{u})$ on both commodity prices and generalized transportation costs suggests that the traditional spatial price equilibrium conditions can be enforced through a carefully constructed $\mathbf{A}(\pi, \mathbf{u})$ matrix.

In light of the preceding notation and the definition of the $\mathbf{A}(\pi, \mathbf{u})$ matrix, the following identity relating the generalized transportation demand vector to activity analysis submatrices holds:

$$\mathbf{T}(\pi, \mathbf{y}, \mathbf{u}) = \Gamma[\mathbf{A}(\pi, \mathbf{u})]^T \mathbf{y} \tag{1}$$

where $\Gamma(\cdot)$ is an operator which properly calculates inter-market (that is, between nodes) generalized network transportation demands for a given activity matrix.

To continue our development of an integrated model, we need to select a routing principle for flows on the detailed generalized transportation network. Specifically, we employ the concept of a Cournot-Nash equilibrium (CNE) based on the arc and path impedance operators used by each subnetwork IN_i to decide distribution and detailed routings. Furthermore, we assume in this streamlined presentation that each path is associated with one and only one subnetwork; this assumption is conceptually easy to relax but requires more involved notation. For each subnetwork-specific path, we assume that there is a *unit* impedance measure

$$\Phi_p^i \left[c_p^i(\mathbf{h}) \right] \tag{2}$$

where $\Phi_p^i[.,.]$ is a monotonic operator that modifies the generalized transportation cost to create the appropriate impedance measure for path p of infrastructure network IN_i.

We are now ready to give a statement of an integrated SCGE model:

DEFINITION 1. *A spatial computable general equilibrium model of perfect competition, SCGE* $(\mathbf{b}, \mathbf{d}, \mathbf{A}, \mathbf{c}(\mathbf{h}))$, *is to find a nonnegative vector* $(\pi^*, \mathbf{y}^*, \mathbf{h}^*, \mathbf{u}^*)$ *such that the following constraints are satisfied:*

1. *no activity in any location earns a positive profit:*

$$-\mathbf{A}(\pi^*, \mathbf{u}^*)^T \pi^* \geqslant 0 \tag{3}$$

2. *an activity in a location which is earning a negative profit is not operated and an operated activity earns zero profit:*

$$[-\mathbf{A}(\pi^*, \mathbf{u}^*)^T \pi^*]^T \mathbf{y}^* = 0 \tag{4}$$

3. *no commodity produced in any location is in excess demand:*

$$\mathbf{b} + \mathbf{A}(\pi^*, \mathbf{u}^*)\mathbf{y}^* - \mathbf{d}(\pi^*) \geqslant 0 \tag{5}$$

4. *a commodity in excess supply is free and a positive price implies market clearing via Walras' Law:*

$$[\mathbf{b} + \mathbf{A}(\pi^*, \mathbf{u}^*)\mathbf{y}^* - \mathbf{d}(\pi^*)]^T \pi^* = 0 \tag{6}$$

5. *excess path costs are nonnegative:*

$$\mathbf{c}(\mathbf{h}^*) - \Lambda^T \mathbf{u}^* \geqslant 0 \tag{7}$$

6. *utilized paths have zero excess costs and paths with positive excess costs are not used:*

$$[\mathbf{c}(\mathbf{h}^*) - \Lambda^T \mathbf{u}^*]^T \mathbf{h}^* = 0 \tag{8}$$

7. *generalized transportation flows are conserved:*

$$\Lambda \mathbf{h}^* - \Gamma[\mathbf{A}(\pi^*, \mathbf{u}^*)]^T \mathbf{y}^* = 0 \tag{9}$$

The definitions

$$\mathbf{x} \equiv \begin{bmatrix} \pi \\ \mathbf{y} \\ \mathbf{h} \\ \mathbf{u} \end{bmatrix} \tag{10}$$

$$\mathbf{H}(\mathbf{x}) \equiv \begin{bmatrix} \mathbf{b} + \mathbf{A}(\pi, \mathbf{u})\mathbf{y} - \mathbf{d}(\pi) \\ -\mathbf{A}(\pi, \mathbf{u})^T \pi \\ \mathbf{c}(\mathbf{h}) - \Lambda^{\mathrm{T}}\mathbf{u} \\ \Lambda\mathbf{h} - \Gamma[\mathbf{A}(\pi, \mathbf{u})]^T \mathbf{y} \end{bmatrix} \tag{11}$$

lead us to a concise nonlinear complementarity formulation of the integrated model defined by (3)-(9):

$$\left.\begin{array}{r} \mathbf{H}(\mathbf{x})^T \mathbf{x} = 0 \\ \mathbf{H}(\mathbf{x}) \geqslant 0 \\ \mathbf{x} \geqslant 0 \end{array}\right\} \quad \mathrm{NCP}(\mathbf{H}, \mathbf{x}), \tag{12}$$

where

$$\mathbf{x} \equiv \{(\pi,\ \mathbf{y},\ \mathbf{h},\ \mathbf{u}) : \pi \in \Re_+^{|\mathbf{R}|\times|\mathbf{N}|},\ \mathbf{y} \in \Re_+^{|\mathbf{R}|\times|\mathbf{M}|},\ \mathbf{h} \in \Re_+^{|\mathbf{P}|\times|\mathbf{N}|},\ \mathbf{u}\} \tag{13}$$

with

$$\mathbf{u} \equiv \{u_{rs}^i : u_{rs}^i = \inf_{p \in \mathbf{P}_{rs}} [c_p^i(\mathbf{h})]\ \forall\, r, s \in \mathbf{R}, i \in \mathbf{N}\} \in \Re_{++}^{|\mathbf{R}|\times|\mathbf{R}|\times|\mathbf{N}|} \tag{14}$$

Note that in (14), $\mathbf{u} \in \Re_{++}^{|\mathbf{R}|\times|\mathbf{R}|\times|\mathbf{N}|}$ denotes that $\mathbf{u}$ is defined in the positive half plane of the corresponding Euclidean space. This condition is necessary for NCP($\mathbf{H}$, Φ) to be equivalent to conditions (3)–(9) and is ensured by positive generalized network transportation arc costs which compel path costs to also be positive. Specifically, the cost positivity assumption in (14) is necessary for the fulfillment of (9).[1] Positive generalized transportation costs are not problematic in a spatial model of real physical and economic processes.

The so-called conservation constraint (9) is really a result of combining the traditional conservation constraints and the inter-locational generalized transportation demand equation (1). As such (9) provides the critical *linkage or*

[1] Briefly, expression (8) ensures that $\mathbf{u}^*$ is a vector of minimum costs. So, if transportation costs are positive it must be that each $u_{rs}^{i*} > 0$, as stated in (14). Since in NCP($\mathbf{H}$, Φ) we have $[\Lambda\mathbf{h} - \Gamma[\mathbf{A}(\pi, \mathbf{u})]^T \mathbf{y}]^T \mathbf{u} = 0$, it follows that if each $u_{rs}^{i*} > 0$ then (9) holds always.

coupling constraints between the CGE and the CNE. Although, as presented, (9) is a relatively straight forward accounting identity first proposed by Friesz et al. (1998) for inclusion in SCGE calculations to ensure submodel consistency. It will be noted that (9) is, in the most general case, a nonlinear equality constraint, holding the potential of making the SCGE model nonconvex even if the generalized transportation cost (congestion) functions are monotonically increasing. Yet, if such a nonconvexity arises it is for intrinsic reasons, since (9) is an indisputable accounting identity and *not* an approximation or modeling assumption.

The way this simple model is formulated implies that generalized transportation networks are provided by an external agent, and, thus, there is potentially a leakage of income out of the economy as it is modeled here. This may be matched by an internal transportation sector devoted to physical movements of goods and services in the activity matrix which demands local resources and exports transportation services. However, in the simple model of this section there is no guarantee of balance in the trade of transportation services. Such a balance could, however, be enforced through the use of additional constraints which are omitted here to simplify the mathematical formulation and emphasize the simultaneous and consistent calculation of economic activities and detailed network flows, which has not been heretofore possible. That is to say, the framework presented here with a consistent treatment of economic activities and detailed network flows can serve as a point of departure for a wide variety of related SCGE models with increased realism that are based on multi-layer generalized transportation networks.

Rather evidently, variational inequality formulations of the above SCGE $(\mathbf{b}, \mathbf{d}, \mathbf{A}, \mathbf{c}(\mathbf{h}))$ problem are also possible, allowing one to use existing results on variational inequalities to establish existence properties and construct efficient algorithms to compute solutions of SCGE models. An existence result for this class of models based on the ideas of Goldsman and Harker (1990) may be easily stated.

7.2 Multi-Layer Dynamic Infrastructure Network Models

If the layers are viewed as the collective engineering means by which a market economy completes its transactions, one is led to a *spatial computable general equilibrium* (SCGE) model of INs. In a SCGE model, generalized transportation networks are represented in the considerable detail needed for engineering analyses, rather than in the aggregate fashion typical of economic forecasting models. This detailed representation of the INs enables one to study the influence of specific IN features on all economic sectors and all locations through *comparative statics*. Equilibria computed with an SCGE model, may in turn be used to construct a disequilibrium dynamic model of coupled INs.

Such a dynamic IN model allows study of nonlinear synergies and catastrophes among infrastructure technologies that would go unnoticed so long as the traditional one-network-at-a-time paradigm is employed.

We discuss below how both the equilibrium and the dynamic multilayer IN models can be coupled to traditional engineering notions of capital budgeting and network design to create a new generation of IN mathematical and software tools. These mathematical and software tools are intended to provide a foundation for new advanced engineering decision support systems for IN planning and design. The discussion that follows is rather technical; someone wishing to gain a highlevel overview of our thinking can read the prose remarks introducing each section and subsection, without bothering with the mathematical symbology.

7.3 Inter-Layer Coupling Constraints

We reiterate that , although a considerable amount of work has been done on SCGE models, little attention has been given to creating a multi-layer network spatial computable general equilibrium model that recognizes the coupled nature of individual INs. It is clear that such a multi-layer model will require the consideration of inter-layer coupling constraints. While the exact nature of these constraints is yet to be determined, it is clear that the layers are not independent. Indeed, as discussed in the introduction, most of the interesting public policy questions in this area arise because of the inter-layer dependencies. Hence, considerable attention will be devoted to these issues.

In the context of coupling constraints, it is important to consider cooperative, non-cooperative and collusive models. The non-cooperative case can be viewed both as a Cournot-Nash mathematical game and as a Stackelberg-Cournot-Nash mathematical game. Although it is too early to be definitive about how best to mathematically describe imperfect, collusive network competition, we are inclined at this time to recommend a hierarchical mathematical program with side constraints to describe relevant market regulations as well as the core of the game being played.

7.4 Dynamics of Disequilibrium

SCGE models like that of Section 7.1 may be employed to calculate a general equilibrium among spatially separated markets that will describe the detailed steady state flows on INs. We believe that in a historical period of rapid information technology change, like the present, it is insufficient to calculate equilibria and to base engineering plans and designs on those equilibria. Rather, it is essential to determine the time-varying flows on INs, as these include transient phenomena that may make the difference between the success or failure of infrastructure and network engineering projects. An example is provided by the current California energy crisis where planning based on steady state prices has been wholly inadequate.

What is needed are dynamics whose steady states are the equilibria that result from SCGE models like that of the previous section. There are a variety of approaches that may be followed in constructing such *equilibrium tending IN dynamics*. Two approaches that we intend to consider are:

1. *Equilibrium-tending adjustment mechanisms* that make the time rate of change of any state variable proportional to the "distance" of that variable from its equilibrium value. This class of dynamics has recently been emphasized by Fujita et al. (1999) to describe the formation of cities viewed as concentrations of population and infrastructure; and
2. *Projective dynamics* that exploit the concept of a minimum norm projection operator to imbed the equilibrium fixed point problem that is the SCGE model into the definition of the time rate of change of a state variable. See Smith et al. (1997) for a review and comparison of different classes of projective dynamics and illustrations of how they are constructed from a known equilibrium model.

7.4.1 Equilibrium Tending Dynamics

Let us suppose that the multi-layer infrastructure network SCGE model has been solved and that its solution $\mathbf{z}^* = (\pi^*, \mathbf{y}^*, \mathbf{h}^*, \mathbf{u}^*)$ includes the vectors of equilibrium commodity prices π^*, equilibrium production and consumption quantities $\mathbf{y}^*$, equilibrium path flows $\mathbf{h}^*$ and equilibrium generalized transportation costs $\mathbf{u}^*$. The time rates of change of the state variables $\mathbf{z}$ may be modeled as proportional to the "distance" from equilibrium:

$$\frac{d\mathbf{z}}{dt} = \zeta \mathbf{F}\left(\|\mathbf{z} - \mathbf{z}^*\|\right) \tag{15}$$

where $\|.\|$ denotes the norm (or distance), ζ is a vector of proportionality constants and $\mathbf{F}$ is a monotonic transformation.

Furthermore, side constraints can be added to these dynamics to form a richer model. In the event, as will be the case for capital budgeting, these dynamics are used to form an optimal control model, the side constraints can be stated as explicit control, state, or mixed control-state constraints. When the dynamics are used alone to describe the IN evolution for given initial conditions, one must perform the necessary algebra to embed the constraints in the right hand sides; certainly this is the approach taken by Fujita et al. (1999). Sometimes this approach of direct substitution is not practical because the side constraints cannot be manipulated in a way that allows such a substitution to occur in closed form. In this eventuality, an alternative approach based on the minimum norm projection operator is needed in order to create numerically tractable constrained dynamics.

7.4.2 Minimum Norm Projective Dynamics

In spatially separated, multi-commodity markets of the kind being considered here, there is no particular reason to believe that the rate at which the market moves towards an equilibrium is proportional to the distance from that equilibrium. Indeed, some profit opportunities may be easy to respond to even though they are small (e.g., the difference in the price of gasoline at nearby stations) and some profit opportunities may be difficult to respond to even though they are large (e.g., excess demand for Internet services). Hence, one is led to consider other types of dynamics.

One alternative approach that has received some attention is *projective dynamics*. The idea is to start by describing an adjustment mechanism that its behaviorally appealing (e.g., the relevant rate of change depends on the cost of the underlying infrastructure, or the rate of change depends on the existing capacity in the infrastructure network). Unfortunately, by themselves, these kinds of adjustment mechanisms may lead to infeasibilities. For example, a large difference in gasoline prices between two stations on Saturday, in and of itself, might seem to warrant a large change in the number of cars that use each station on Sunday. However, the number of people that purchase gasoline on that Sunday might not be as large as the number of people that "should" switch. One way to correct these kinds of infeasibilities is to use a simple adjustment mechanism and then *project* the outcome onto the feasible region using the *minimum norm projection*. Indeed, the so-called "cobweb" model can be thought of as an example of projective dynamics. The primary advantage of projective dynamics is that they allow one to construct descriptive models with side constraints that are now amenable to direct algebraic incorporation into the dynamics.

Using the notation of the previous discussion on equilibrium-tending dynamics, the essence of projective dynamics is captured by noting that

$$\frac{d\mathbf{z}}{dt} = \lim_{\Delta t \longrightarrow 0} \frac{\mathbf{z}(t + \Delta t) - \mathbf{z}(t)}{\Delta t} \tag{16}$$

and modeling the future state according to

$$\mathbf{z}(t + \Delta t) = \mathcal{P}_{\Omega}[\mathbf{z}(t) + \rho \Delta t \mathbf{S}(t)] \tag{17}$$

where $\mathbf{S}(t)$ is a signal (sometimes called the "velocity") that determines the increment/decrement of the current state to form the new state and $\mathcal{P}_{\Omega}[.]$ denotes the minimum norm projection of its argument onto a set of constraints. The set Ω is comprised of all pertinent technological, budgetary, policy and regulatory constraints. Growing experience with such dynamics [see Smith et al. (1997)] indicates that projective dynamics have substantial explanatory power and are valid representations of many complex socio-technical systems.

7.5 Network Design Criteria

In order to design INs and to allocate funds for their operations, maintenance and replacement, one must articulate investment criteria. The one we shall emphasize is the net present value of benefits. In calculating this present value one may measure gross benefits as the sum of consumers' and producers' surpluses. Furthermore, costs will be measured as actual economic costs *specifically including congestion costs for each IN layer considered*. Naturally, net benefits are determined by subtracting such generalized costs from gross benefits. It is important to recognize that there is a *fundamental measurement problem that has prevented rigorous net benefit calculations of this sort from being made heretofore for networks with elastic user demands*. This is the difficulty surrounding the consumers' surplus line integral. Because of the immense importance of this problem to multi-layer IN capital budgeting, we must give it a fairly detailed description.

All previous IN design and capital budgeting models reported in the literature deal either with constant (inelastic) user demand or presume elastic demand functions are separable. Yang and Bell (1998) and Huang and Bell (1998) are among the few to have considered network design in the presence of elastic demand, and they employ separable demand functions. However, when consumption alternatives exist for network users, the assumption of demand separability or the assumption of fixed demands is completely inadequate. Moreover, for the price-quantity pair $(\Theta(\mathbf{T}), \mathbf{T})$ consumers' surplus for infrastructure services is given by the line integral.

$$CS(\mathbf{T}) = \sum_{w=1}^{q} \left[\oint_{\mathbf{0}}^{\mathbf{T}} \Theta_w(\mathbf{v})\, dv_w - \Theta_w(\mathbf{T})\, T_w \right] \tag{18}$$

which includes the separable case.[2] The first term of (18) measures gross economic benefits and the second term is the payment made for benefits in terms of congestion costs. Note that a line integral *must* be employed in order to give an exact representation of net economic benefits in the presence of elastic, nonseparable demand functions. This is because consumers' surplus necessarily involves the integration of functions of several variables when demand functions are not separable. This is problematic because it is well known that a line integral does not have an unique, unambiguous value unless the Jacobian matrix formed from its integrand is symmetric. Such symmetry restrictions amount to a requirement that cross price elasticities of demand be proportional to one another, which is unlikely in any real world setting. See

[2] In (18) $\mathbf{u} = \Theta(\mathbf{T})$ is the relevant inverse demand function that specifies IN service price for a given origin-destination flow $\mathbf{T}$ on the IN of interest.

Jara-Diaz and Friesz (1982) for a discussion of these subtleties. Note also that this symmetry restriction arises in all static settings regardless of how the equilibrium problem is formulated because it stems from a theoretically rigorous statement of user net benefits in the presence of elastic demand; that is, from the inclusion of a rigorous statement of consumers' surplus. As such, the static equilibrium constrained network capital budgeting problem with elastic service demands *cannot be solved* since we have no guidance about what path of integration to employ in evaluating the consumers' surplus line integral intrinsic to its formulation.

However, the dynamic models discussed above may be used to calculate consumers' surplus in a mathematically rigorous fashion. It turns out that this is quite easy when adjustment dynamics like (15) or alternatives based on the minimum norm projection are known. In a dynamic setting (18) becomes

$$CS(\mathbf{T}) = \int_0^L \exp(-rt) \sum_{w=1}^{q} \left[\oint_{\mathbf{0}}^{\mathbf{T}} \Theta_w(\mathbf{v})\, dv_w - \Theta_w(\mathbf{T})\, T_w \right] dt$$

where r is a fixed discount rate, L is the length of the planning horizon considered, and t is of course continuous time. Of course a complete consideration of benefits and costs means there will be other integrals to express consumers' surplus for other-than-infrastructure consumption as well as expressions for all relevant producers' surpluses.

7.6 Mathematical Formulation of the Dynamic Multi-layer IN Design Model

Traditionally, optimal network design and capital budgeting have been carried out in a static environment [Friesz (1985)]. The well-known Braess' paradox requires that static design have network equilibrium constraints and that they be articulated as a mathematical program with variational inequality or other equilibrium-enforcing constraints [Friesz et al. (1992)]. Unfortunately, such equilibrium constrained optimization models completely ignore the potential disequilibria that can arise from changes in the *effective capacity* of an IN arc. This static perspective holds the potential of recommending not only suboptimal designs, but also of recommending economically infeasible designs. Such erroneous findings are possible when the present value of disequilibrium impacts is substantially negative, as can occur when the immediate disequilibrium response is a sharp congestion increase or when the disequilibrium response is a mild congestion increase of relatively long duration. An example drawn from current events is provided by a network construction project intended to enhance telecommunications and data capacity through laying fiber optic cable but which produces road traffic congestion for weeks or months prior to its completion; in this case, the present value of construction impacts may be sufficiently negative to eradicate all positive benefits in the post-construction period.

In light of these remarks, it is necessary to articulate a disequilibrium capital budgeting model, by which is meant a model for selecting optimal effective capacity enhancement *trajectories* for network arcs, with the time evolution of flows and perceived path costs described by an appropriate disequilibrium adjustment process. A model of this type will determine a temporal IN capacity enhancement plan which recognizes that capacity perturbations produce disequilibria which adjust toward equilibrium. Such a model is naturally expressed as an optimal control problem or infinite dimensional mathematical program. That is

maximize present value of net benefits

subject to

state dynamics

budget constraints

layer-to-layer coupling constraints

non-negativity constraints

arc capacity upper bound constraints

See Peeta et al. (2005) for the details of such a formulation. For our present discussion it is enough to recognize that a formulation like that above constitutes an optimal control problem. That optimal control problem may be expressed as an infinite dimensional mathematical program in the appropriate function spaces.

7.7 Solution Techniques

Our envisioned multi-layer capital budgeting model is a hierarchical infinite dimensional mathematical program; as such a variety of classical numerical methods are available for its solution: two point boundary value problem shooting methods, dynamic programming, time discretization/finite dimensional mathematical programming and constrained descent methods in infinite dimensional topological vector spaces. However, both the number of variables and the number of constraints needed for the articulation of this model for an actual multi-layer IN will certainly be in the thousands. Matters are further complicated by the presence of explicit path variables in the model's formulation; this is a direct result of its dynamic nature and the fact that IN generalized transportation demands naturally occur at the origin-destination level. A third feature of the multi-layer IN capital budgeting model is that it is unavoidably nonconvex; this nonconvexity arises from the nonlinear equality constraints expressing the coupling of the various network layers.

7.7.1 The Conventional Wisdom for Static Network Design

In computational network modeling many believe that the presence of explicit path variables in a model formulation makes that model numerically intractable because of the potential for extraordinarily large numbers of paths in networks of realistic size. This point of view, which typically implies complete path enumeration, has always been incorrect and only recently has the fear of path variables begun to subside among network model builders. Path variables are handled in all successful algorithms by path generation schemes which only calculate paths as they are needed; this philosophy – which is essentially the technique of column generation well known to mathematical programmers – is at the heart of the well known and successful application of the Frank-Wolfe algorithm to static network equilibrium calculations. What many do not realize is that although the Frank-Wolfe algorithm is usually implemented in a way that discards path information and saves only arc flows, it is straightforward to develop and implement Frank-Wolfe type software which generates and saves path identities. Such software is no less intrinsically computationally efficient than Frank-Wolfe software which saves only arc information; the main additional computational burden is the storage of the chains of arcs defining paths. When the number of paths becomes large enough that some kind of virtual storage is needed to save paths, there can in principle be processing delays when paths are retrieved from memory. Yet even this is not very significant, since paths are typically associated with origin-destination pairs and there are on the order of only ten meaningful paths for each origin-destination pair, a situation which allows the origin-destination pair to serve as a pointer which takes one to more or less the exact memory location instantly. Moreover, Friesz et al. (1992) and Friesz et al. (1993) have shown how the Frank-Wolfe algorithm can be used very effectively as an *a priori* path generator. This is done by making the network increasingly congested through demand increases and saving path information from the Frank-Wolfe algorithm until the generated paths exceed a pre-established circuitousness criterion.

Even though we do not believe the presence of path variables will in and of themselves constitute a major computational impediment, the very large overall size and nonconvex nature of the multi-layer IN capital budgeting model suggest that nontraditional numerical methods for its solution must be explored. We offer next a brief overview of some ideas relevant to developing such nontraditional methods for our coupled IN capital budgeting model.

7.7.2 Simulated Annealing

One approach that has been used successfully for non-cooperative static games of both the Cournot-Nash and Stackelberg type is simulated annealing (SA). It should be equally successful for finite-difference approximations of the continuous time models we have been discussing. SA has its roots in statistical mechanics and is in effect a kind of probabilistic search method. SA was originally developed by Kirkpatrick et al. (1983) to solve combinatorial or discrete optimization problems. Vanderbilt and Louie (1984) showed how this method can be employed to obtain global solutions for continuous optimization problems. In fact, SA has proven to be an effective technique for virtually any continuous nonconvex optimization problem when rapid response times are not mandated by the decision environment; see Anandalingam (1989). In SA, as well as in more recent AI algorithms for mathematical programming based on neural networks and genetic analogies, there are a lot of parameters under the control of the analyst. This gives the analyst freedom to tailor the AI algorithm of choice to a particular application. Sometimes, however, there is a price to be paid for this extra freedom; namely, in some problems the selection of the various parameters can initially pose a serious challenge, making convergence difficult to achieve. This drawback, however, needs to be weighed against the relative simplicity of AI software code for solving optimization problems. In fact, our experience in training students and professionals indicates that most individuals with some prior experience with one of the common programming languages can write adequate SA software code after only a few hours of exposure. This contrasts sharply with sophisticated decomposition algorithms and other techniques traditionally favored by mathematical programmers for large network problems that can require substantial mathematical sophistication and careful software engineering to fully understand and implement.

For a description of the details of applying SA to mathematical formulations like the one suggested here, see Friesz et al. (1992) and Friesz et al. (1993). In fact, previously unknown equilibrium network design solutions which are probably global optima have been found by Friesz et al. (1992) using this approach for the much studied Sioux-Falls single infrastructure road transportation network. The numerical results reported by Friesz et al. (1992) and Friesz et al. (1993) demonstrate unequivocally that simulated annealing is a viable and effective algorithmic approach to network capacity expansion in an equilibrium setting.

7.7.3 Agent-Based Simulation

Peeta et al. (2005) argue that dynamic multi-level network design may be dealt with effectively by re-expressing the models discussed above via agent-based simulation (ABS). They present numerical examples that show the ABS perspective is successful. At the present moment only the ABS-based approach to solution of large-scale dynamic infrastructure network design models has been applied to obtain numerical results for the Stackelberg-game formulations proposed in this paper. Although simulated annealing has been successful for static Stackelberg games, no application of that technique to dynamic Stackelberg models of infrastructure networks has been carried out.

References

Albert, R., H, Jeong and A.-L. Barabási et al. (1999): Diameter of the World-Wide Web, Nature, 401, 130–131.

Albert, R., and Barabasi, A.-L. (2002): Statistical Mechanics of Complex Networks Reviews of Modern Physics, 74, 47–94.

Anandalingam, G. (1989): Simulated Annealing and Resource Location in Computer Neworks, Simulation Conference Proceedings, 1989. Winter, 980–988.

Crossley, W. A. (2004): System of Systems: An Introduction of Purdue University Schools of Engineering's Signature Area, Engineering Systems Symposium, MIT Engineering Systems Division, Cambridge, MA, http://esd.mit.edu/symposium/pdfs/papers/crossley.pdf

Friesz, T. L. (1985): Transportation Network Equilibrium, Design and Aggregation, Transportation Research, 19A, 413–428.

Friesz, T. L., H. J. Cho, N. J. Mehta, R. L. Tobin and G. Anandalingam (1992): A Simulated Annealing Approach to the Network Design Problem with Variational Inequality Constraints, Transportation Science, 26, 18–26.

Friesz, T. L., D. Bernstein, T. E. Smith, R. L. Tobin and B. W. Wie (1993): A Variational Inequality Formulation of the Dynamic Network User Equilibrium Problem, Operations Research, 41, 179–191.

Friesz, T. L., Z.-G. Suo and D. H. Bernstein (1998): A Dynamic Disequilibrium Interregional Commodity Flow Model. Transportation Research, 32B, 457–483.

Fujita, M., P. R. Krugman and A. J. Venables (1999): The Spatial Economy: Cities, Regions, and International Trade, MIT Press.

Goldsman, L. and P. T. Harker (1990): A Note on Solving General Equilibrium Problems with Variational Inequality Techniques, Operations Research Letters, 9, 335–339.

Gupta, H. M., and Campanha, J. R. (2003): Is Power Law Scaling a Quantitative Description of Darwin's Theory of Evolution?, arXiv:cond-mat/0311542

Huang, H. J. and M. G. H. Bell (1998): A Study of Logit Assignment Which Excludes All Cyclic Flows, Transportation Research, 32B, 401–412.

Jara-Diaz, S. R. and T. L. Friesz (1982): Measuring the Benefits Derived from a Transportation Investment, Transportation Research, 16, 57–77.

Keating, C., Rogers, R. Unal, R., Dryer, D., Sousa-Poza, A., Safford, R., Peterson, W. and Rabadi, G. (2003). "System of Systems Engineering," Engineering Management Journal, 15 (3), 36–45.

Kirkpatrick, S., C. D. Gelatt and M. P. Vecchi (1983): Optimization by Simulated Annealing, Science, 220, 671–680.

Kretchschmer, H. (1994): Coauthorship Networksof Invisible College and Institutionalized Communities, Scientometrics, 30, 363–369.

Levis, A. H. (2004): "Perspectives in Systems Engineering," excerpt from course lecture Architecture-based Systems Engineering for Senor Leaders, Armed Forces Communications and Electronics Association.

Lotka, A. J. (1926): The Frequency Distribution of Scientific Production, J. Washington Academy of Science, 16, 317–323.

Mathiesen, L. (1985a): Computational Experience in Solving Equilibrium Models by a Sequence of Linear Complementarity Problems, Operations Research, 33, 1225–1250.

Mathiesen, L. (1985b): Computation of Economic Equilibria by a Sequence of Linear Complementarity Problems, Mathematical Programming Study, 23, 144–162.

Nagurney, A. (2006): http://supernet.som.umass.edu/

Newman, E. J. (2003): The Structure and Function of Complex Networks, SIAM Review, 45, 167–256.

Newman, M. (2003): The Structure and Function of Complex Networks, SIAM Review, 45, 167–256.

Parisi, G. (2004): Complex Systems: a Physicist's Viewpoint, arXiv:cond-mat/0205297

Peeta, S., P. Zhang and T. L. Friesz (2005): Dynamic Game Theoretic Model of Multi-Layer Infrastructure Networks, *Networks and Spatial Economics*, 5, 147–178.

Persson, O. and M. Beckmann (1995): Locating the Network of Interacting Authors in Scientific Specialties, Scientometrics, 33, 351–356.

Sage, A. P., and Cuppan, C. D. (2001): "On the Systems Engineering and Management of Systems of Systems and Federations of Systems," Information, Knowledge, Systems Management, 2(4), 325–45.

Scarf, H. E. and T. Hansen (1973): The Computation of Economic Equilibria, Yale University Press.

Sheffi, Y. (1985): Urban Transportation Networks: Equilibrium Analysis With Mathematical Programming Methods, Prentice Hall.

Smith, T. E., T. L. Friesz, D. H. Berstein and Z.-G. Suo (1997): A Comparative Analysis of Two Minimum Complementarity and Variational Problems: State of the Art. SIAM.

Vanderbilt, D. and S. G. Louie (1984): A Monte Carlo Simulated Annealing Approach to Optimization over Continuous Variables, Journal of Computational Physics, 56, 259–271.

Venkatasubramanian, V., Politis, D. N., and Patkar, P. R. (2004): Entropy Maximization as a Holistic Design Principle for Complex Optimal Networks and the Emergence of Power Laws, arXiv:nlin/0408007

Watts, D. J. and S. H. Strogatz (1998): Collective Dynamics of Small World Networks, Nature, 393, 440–442.

Watts, D. J. (2003): Six Degrees: The New Science of the Connected Age, Norton.

Yang, H. and M. G. H. Bell (1998): Models and Algorithms for Road Network Design: A Review and Some New Developments.

Chapter 4

Dynamic Competition on Networks: Differential Variational Inequalities, Limited Warfare and Internet Vulnerability

Terry L. Friesz

Harold and Inge Marcus Chaired Professor of Industrial Engineering,
The Pennsylvania State University, University Park, PA 16802, USA,
E-mail: tfriesz@psu.edu

Abstract In this paper we are concerned with a class of mathematical problems that arise when modelling dynamic competition among conflicting agents that may or may not reach an equilibrium. We call this class of problems *differential variational inequalities*. We are further interested in the application of models belonging to this class of problems to the study of (1) policy constrained, *limited warfare* and (2) *Internet vulnerability*.

Keywords: differential variational inequalities; networks; dynamic games; internet modeling

1. Introduction

The latter half of the twentieth century saw impressive achievements in the modeling, analysis and computation of competitive static equilibria, as is underscored by the joint award of a Noble Prize to John Nash, John Harsanyi and Reinhard Selten in 1993 for their fundamental work on mathematical games and the relationship of games to equilibrium and optimization. Mathematicians, game theorists, operations researchers, economists, biologists and engineers have employed noncooperative mathematical games and the notion of equilibrium to model virtually every kind of competition.

In particular, non-cooperative game theoretic models have been successfully employed to study economic competition in the marketplace, highway and transit traffic in the presence of congestion, wars, and both intra- and inter-species biological competition. One of the key developments that has made such diverse applications practical is our ability to compute static game-theoretic equilibria as the limit of a sequence of well-defined subproblems solvable by variants of traditional nonlinear programming, fixed point, and complementarity algorithms.

In many applications, intermediate disequilibrium states of mathematical games are intrinsically important. When this is the case, disequilibrium adjustment processes[1] must be articulated, thereby, forcing explicit consideration of time. As a consequence, the modeling of competitive disequilibria involves dynamic or differential games. While great progress has been made in modeling and computation of equilibria or steady states of competitive systems, game-theoretic disequilibria and moving equilibria[2] are relatively uninvestigated by comparison.

The main body of technical literature relevant to game-theoretic disequilibria and moving equilibria is that pertaining to so-called *differential games*, a field of inquiry widely held to have been originated by Isaacs (1965). Although a rather substantial body of literature known as *dynamic game theory* has evolved from the work of Isaacs (1965), that literature continues to be strongly influenced by the emphasis of Isaacs on the relationship of such games to dynamic programming and to the Hamilton-Jacobi-Bellman partial differential equation.[3] A consequence of this classical point of view is that full use of the mathematical apparatus of variational inequalities, discovered originally in the context of certain free boundary value problems in mathematical physics, has not occurred in the study of dynamic games. By contrast, in the last fifteen years, variational inequalities have become the formalism of choice for applied game theorists and computational economists solving various static equilibrium models of competition. The "hole" in the dynamic game theory literature owing to this failure to fully exploit the variational inequality perspective is significant, for variational inequalities substantially simplify the study of existence and uniqueness. A variational inequality perspective for infinite dimensional dynamic games also leads directly to function space equivalents of the standard finite dimensional algorithmic philosophies of feasible direction and projection familiar from nonlinear programming. Because our point of view in this paper departs from that of differential game theory and classical

[1] Such adjustment processes are typically expressed as difference equations or differential equations. Sometimes, it is necessary to use a mixture of both types of dynamics; the adjustment processes are then referred to as *differential-difference equations*.

[2] We define *moving equilibrium* in a subsequent section; however, it is easiest at this juncture to think of a moving equilibrium as a trajectory of decision variables that maintains the same "balance" among those variables throughout time, although the variables themselves are time-varying.

[3] See for example Basar and Olsder (1998).

dynamic game theory, we refer to the class of problems emphasized herein as *differential variational inequalities*, terminology we shall shortly formally define in Section 5.

2. Scope of this Paper

In mathematical modeling of competition among conflicting agents, regardless of whether the application is one drawn from biology, economics or military operations, there are three fundamental approaches:

1. statistical inference based on historical data,
2. detailed microsimulation, or
3. mathematical game theory,

where *mathematical game theory* is understood to be the process of constructing a system of equations, inequalities and/or extremal principles that describe the behavioral aspects of the competitive system of interest and technological considerations affecting that system. This mathematical description is necessarily an approximation of the real system, and it may be subject to mathematical analysis to uncover qualitative properties (existence, stability, uniqueness, and decision rules) and to numerical analysis to determine solutions needed for prediction and control in specific decision environments. Generally speaking, each of the three main categories of modeling approaches suffers from certain weaknesses and enjoys certain strengths. In particular, statistical inference – since it is based on historical data – cannot reliably predict system states when significant structural changes occur within the system of interest. Furthermore, microsimulation is generally very labor intensive and tends to involve numerous *ad hoc* assumptions. Because of the fact that linkages among subsystems of a microsimulation are usually accomplished through "if/then" logic, the resulting mathematical model is non-differentiable. As a consequence, qualitative analyses of microsimulation models are extremely difficult.

Our presentation in this paper hinges on the distinction between microsimulation models and a class of mathematical game theoretic models that may be called *game-theoretic screening models*. This distinction is most easily made by contrasting extreme versions of microsimulation and screening models. To this end, consider an extremely detailed microsimulation that is virtually an exact replication of the decision processes of the competitive agents of interest and which outputs an agent-by-agent, decision-by-decision, minute-by-minute description of the competitive system of interest.[4] It is easy to

[4] In fact such detailed microsimulations have been proposed and implemented by Los Alamos Laboratory to describe human activity patterns for the entire United States. The Los Alamos microsimulation requires the use of powerful, mainframe supercomputers.

understand that such a detailed microsimulation consumes staggering amounts of computational resources and that its software is intrinsically expensive to program and maintain. In contrast is the *screening model* that identifies the key relationships and provides an approximate mathematical articulation of the system of interest using a few state and control variables, an extremal principal and a few key constraints. It is virtually certain that such a screening model will be much less labor intensive to create and will involve substantially fewer computational resources; it may also be less accurate. Of course, in reality, models exist along a continuum that has the two extremes just cited as its end points, and all modeling applications involve a trade-off between the expense of building a model and the accuracy of its predictions. These considerations not withstanding, if a screening model can be put on a solid behavioral foundation, it can be phenomenally cost effective.

It is our thesis in this paper that there is a new perspective for game-theoretic mathematical modeling that results in especially powerful screening models generally applicable to the study and control of competitive dynamic systems and especially relevant to

1. extensions of classical dynamic game theory to more general decision making environments;
2. the general problem of network interdiction and vulnerability; with special emphasis on the Internet; and
3. the modeling of limited warfare in foreign economic environments.

The new perspective on game-theoretic mathematical modeling that we emphasize is called the *differential variational inequality* (DVI).

A DVI is in effect a controlled variational inequality. As such, study of the DVI joins two important and already established branches of research in applied mathematics and operations research, namely variational inequality problems (VIPs) and optimal control problems (OCPs). As we discuss below, DVIs arise naturally in the modeling of dynamic hierarchical competitive systems and provide a means for integrating prescriptive (optimizing) and descriptive (predictive) mathematical models of competition into a single coherent and self consistent mathematical form. Particularly important to the versatility of DVI models of competition is the fact that DVIs include dynamic systems in the form of differential-difference equations among their constraints; these dynamics can either model a moving equilibrium or a disequilibrium adjustment process that may never reach equilibrium.

Moreover, because of the intimate connection of the DVI to the theory of variational inequalities developed for mathematical programming applications, both finite difference and continuous time numerical methods familiar from mathematical programming immediately suggest themselves for obtaining numerical results. We also argue below that the DVI perspective provides a unifying framework for dynamic games, differential games and mathematical programs in function spaces. This unification will almost certainly allow the sharpening of classical existence and uniqueness results for differential and dynamic games; it will also provide a theoretical foundation for algorithms based on sequences of extremal problems even when the game of interest is itself not an extremal problem.

Furthermore, by exploiting the dynamic equations intrinsic to the constraints of a DVI one of the most vexing problems associated with the definition of extremal versions of static noncooperative games may be overcome. We are referring to the well-known result that requires the use of a line integral to represent a general asymmetric noncooperative static game in extremal form. Since line integrals are not generally single valued, the objective function of an extremal formulation of a general asymmetric noncooperative static game is not well defined. It is this property that makes it necessary to use infinite sequences of mathematical programs to solve such static games. By contrast, for dynamic games employing the DVI perspective, we may exploit the dynamic equations among their constraints to define an unambiguous path of integration that makes the extremal problem objective function well defined. We discuss this important feature of DVIs in greater detail in Section 5.2.

In summary then, we believe and do argue below that DVIs extend traditional notions of dynamic and differential mathematical games used in studying competition by providing a mathematical structure that

1. allows the description of systems in disequilibrium that may never reach an equilibrium;
2. greatly facilitates qualitative (existence, uniqueness and stability) analyses and algorithm development;
3. allows the introduction of additional doctrinal, regulatory, economic and technological constraints without analytical or numerical complications; and
4. is unimpacted by symmetry restrictions and does not presume an equivalent extremal formulation.

Each of these features is discussed in separate sections below.

For the sake of brevity, the discussion that follows emphasizes deterministic DVIs. However, because a DVI involves explicit state dynamics, it is possible to extend the notion of a DVI to a stochastic setting involving stochastic differential equations. In fact a stochastic DVI is needed to properly model the Internet, as we later explain in our discussion of Internet vulnerability in Section 7.

3. Variational Inequalities Defined

VIPs or the related format of the nonlinear complementarity problem (NCP), generalize the optimality conditions for nonlinear programs as well as provide the structure for numerous problems in traffic equilibria, economic equilibria, robotics, computational game theory and other areas; see Ferris and Pang (1997) for a survey of such applications. The VIP can be succinctly stated as follows:

DEFINITION 1 (Finite dimensional variational inequality problem). *Let X be a nonempty subset of $\Re^n$ and let $F : X \subseteq \Re^n \to \Re^n$. $VIP(X, F)$ is to find a vector $x^* \in X$ such that the following conditions hold*

$$F\left(x^*\right)^T \left(y - x^*\right) \geqslant 0 \quad \forall y \in X \subseteq \Re^n$$

The VIP is closely related to the nonlinear complementarity problem (NCP). Suffice it to say that under appropriate regularity conditions VIPs and NCPs are equivalent to one another. Under more stringent conditions, the solution of a VIP is the solution of an NCP.

It is of course also possible to articulate the variational inequality problem for an appropriately defined function space. For our purposes, we limit our attention to a non-specific, abstract Hilbert space $\mathcal{H}$ so that the following definition obtains:

DEFINITION 2 (Infinite dimensional variational inequality problem). *Let X be a nonempty subset of the Hilbert space $\mathcal{H}$ and let $F : X \subseteq \mathcal{H} \to \mathcal{H}$. $VIP(X, F)$ is to find a vector $x^* \in \mathcal{H}$ such that the following conditions hold*

$$\left\langle F\left(x^*\right), \left(y - x^*\right)\right\rangle \geqslant 0 \quad \forall y \in X \subseteq \mathcal{H}$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product in $\mathcal{H}$.

See Harker and Pang (1990) for a discussion of existence and uniqueness of solutions as well as algorithms for solving the finite dimensional VIP. See Friedman (1982) for a detailed presentation of infinite dimensional variational inequalities in function spaces. A parallel literature exists for complementarity problems but a review of it is omitted here for the sake of brevity.

For recent works related to specific VIP and NCP algorithms and applications in transportation and energy, see Gabriel and Pang (1992); Pang and Gabriel (1993); Gabriel and Pang (1994); Bernstein and Gabriel (1997); Gabriel and More' (1997); Gabriel and Bernstein (1997); Gabriel (1998ab);

Gabriel and Bernstein (2000); Gabriel, Kydes, and Whitman (2001). Algorithms and applications for infinite dimensional VIPs are discussed by Kinderlehrer and Stampacchia (1980), Friedman (1982), Baiocchi and Capelo (1984), and Le and Schmitt (1997). Interestingly, some of the most recent applications of infinite dimensional variational inequalities, as well as research on numerical methods for their solution, has occurred in the context of option pricing and various exotic securities used in financial markets; see in particular Wilmott et al. (1993).

4. The DVI Concept

The DVI is a special kind of variational inequality in function spaces that has a mathematical structure similar to that of an optimal control problem in that there are readily identifiable state and control variables, state dynamics, and state/control constraints within the feasible region X of the preceding definition of a VIP in Hilbert space. However, unlike optimal control problems, a DVI employs no optimization criterion and there is no *a priori* presumption that an equivalent optimization problem exists; although we shall explain below how such an equivalent problem can be stated for certain plausible and not terribly restrictive regularity conditions that do not include the symmetry requirements familiar from static VIPs.

For the DVI, as for any variational inequality, the optimizing behaviors of individual agents are described by separate necessary conditions that can be collectively enforced by variational inequalities. Furthermore, there is no presumption that individual agents have identical or even similar behaviors; that is, each agent may be described by a separate variational inequality. Moreover, the constraints necessary to modeling the core of any subgame of interest are readily included in the DVI constraint set; as a consequence the full spectrum of games ranging from total collusion through mixed collusion/noncooperation to total noncooperation may be modelled, although pure cooperation and pure noncooperation formulations will remain the most computationally tractable, as is familiar from our experience with finite dimensional variational inequalities.

The previously cited properties of DVIs mean that the DVI structure is especially well suited for the study of general dynamic, hierarchical, multi-agent, multi-behavior games that impose *no a priori symmetry or separability assumptions*. Previous attempts to model such general games required either that the games be written in extensive form (with accompanying computational difficulties) or that their outcomes be found from microsimulation based on *ad hoc* decision rules, since no tractable canonical form for such general games that is amenable to large scale computation has entered the literature.

Such general games are of substantial practical importance as we explain in Section 6, after we introduce some alternative forms of the DVI appropriate to abstract noncooperative Cournot-Nash-Bertrand and Stackelburg-Cournot-Nash-Bertrand dynamic games.

5. DVI Extensions of Classical Noncooperative Mathematical Games

As we have suggested above, the notion of a DVI may be used to extend classical mathematical game theory to a dynamic or moving equilibrium[5] setting without the assumption of symmetry of the Jacobian matrix formed from the performance functions of game agents. Such an extension of classical results is quite important since it allows the immediate generalization to a dynamic setting of the well known Cournot-Nash-Bertrand (CNB) and Stackelberg-Cournot-Nash-Bertrand (SCNB) game theoretical models for static equilibria. Although dynamic CNB and SCNB games have been analyzed heretofore, our reading of the literature indicates that in many instances reported dynamic CNB and SCNB models either invoke notions of symmetry that are unrealistic or employ algorithms that are inefficient in the sense that they do not actively exploit the path of integration intrinsic to the state dynamics.

In order to continue our discussion, it is important to define what is meant by a "dynamic equilibrium", as the words "dynamic" and "equilibrium" are sometimes viewed as mutually exclusive. For our purposes in this paper, we will view a *dynamic equilibrium* as a circumstance wherein equilibrium is enforced at each instant of time although state and control variables will generally be time varying. These variations with respect to time are exactly those needed to maintain the balance of behavioral and economic circumstances defining the equilibrium of interest. Samuelson (1947) has referred to this type of equilibrium as a *moving equilibrium*, a name which is much more descriptive and not as likely to be misunderstood. A moving equilibrium is similar to the usual notion of a static game theoretic equilibrium, except that variables move in unison to maintain equilibrium at each moment of time although there is no steady state.[6]

[5] The phrase "dynamic equilibrium" appears at first blush to be inconsistent since "equilibrium" can connote a steady state; for this reason we prefer the terminology "moving equilibium" defined below.

[6] Of course a formal mathematical definition of a moving equilibrium may be given, a task we avoid here for the sake of brevity. See in particular Samuelson (1947), page 321.

5.1 Formulation of Non-Cooperative Dynamic Equilibrium

Let us consider the case of a single class of game theoretic agents, each of whom has qualitatively identical dynamics and dis-utilities. Moreover, when the relationship among these agents is that of a dynamic Cournot-Nash noncooperative equilibrium, the DVI of interest is described in the following theorem [see Friesz et al. (2001a)]:

THEOREM 3. *A dynamic CNB game-theoretic moving equilibrium among agents described by the vector-valued disutility function $F(u) \in \mathcal{H}$ is equivalent to the following differential variational inequality: find $u^* \in U \subseteq \mathcal{H}$ such that*

$$\left\langle F\left(x^*, u^*, t\right), \left(u - u^*\right)\right\rangle = \int_0^T F\left(x^*, u^*, t\right)\left(u - u^*\right) dt \geqslant 0 \quad \forall u \in U \subseteq \mathcal{H} \tag{1}$$

where $\mathcal{H}$ is a Hilbert space and

$$\langle \cdot, \cdot \rangle \text{ denotes an inner product in } \mathcal{H} \tag{2}$$

$$u \doteq \left(u_i(t)\right)_{i=1}^m \tag{3}$$

$$U = \left\{u : \frac{dx}{dt} = f(x, u, t), \Gamma(u, x, t) \leqslant 0, x(0) = x^0\right\} \tag{4}$$

$$F : \mathcal{V} \times U \times \Re^1 \longrightarrow \mathcal{H} \tag{5}$$

$$f : \mathcal{V} \times U \times \Re^1 \longrightarrow \mathcal{H} \tag{6}$$

$$\Gamma : \mathcal{V} \times U \times \Re^1 \longrightarrow \mathcal{H} \tag{7}$$

for $F(\cdot,\cdot,\cdot)$, $f(\cdot,\cdot,\cdot)$ and $\Gamma(\cdot,\cdot,\cdot)$ convex in (x, u) and weakly continuous on U, and $x(u)$ a mapping from U to $\mathcal{V}$, also a Hilbert space.

We take as implicit that the integral in (1) is a Lesbegue integral and that the dynamics of U, namely

$$\frac{dx}{dt} = f(x, u, t), \tag{8}$$

are flow balance statements that preserve equilibrium at each instant of time. Thus, problem (1) is in effect a differential game of the Cournot-Nash-Bertrand variety, whose solution is a so-called *moving equilibrium*. Note also that this formulation makes no assumptions with regard to differentiability of $F(\cdot,\cdot,\cdot)$ or u. In fact u may and will exhibit jump discontinuities. Furthermore no symmetry restrictions are imposed on $F(\cdot,\cdot,\cdot)$. As a consequence of this generality, DVIs of this type may be used to model a very broad range of phenomena noncooperative, especially noncooperative conflicts involving players with equal quality of information; this includes, but is not limited to, certain types of military warfare and ecommercial Internet driven economic competition.

Because (1) is in the form of a variational inequality we can employ powerful methods from numerical functional analysis for its solution and do not need to develop detailed event based simulation software. That is to say, (1)–(7) allows relatively cheap modeling of dynamic conflicts involving agents that are equal in terms of their information and communication technologies.

A very interesting generalization of (1)–(7) that relaxes somewhat the assumption of the equality of information and communication is to allow the function $F(\cdot,\cdot,\cdot)$ to be replaced by an operator defined on a specific Hilbert space. In particular one may employ operators that reflect the ability of certain agents to look forward in time, without destroying the Cournot-Nash structure that makes computing with (1)–(7) so tractable. These operators for selected agents $i \in S \subset [1, 2, ..., n]$ are such that we make replacements in (1) according to

$$F_i(x, u, t) \longrightarrow F_i(X, U, t)$$

where it is understood that X and U are time histories of the states and controls and include *future* as well as past information; that is

$$\left.\begin{array}{l} X = (x(t) : t \in [0, T]) \\ U = (u(t) : t \in [0, T]) \end{array}\right\} \tag{9}$$

for an especially omniscient agent, considering that $T \in \Re^1_{++}$ is the event horizon. For more ordinary agents the prospective aspect of the look into the future may be limited; that is

$$\left.\begin{array}{l} X = (x(t) : t \in [0, A]) \\ U = (u(t) : t \in [0, A]) \end{array}\right\} \tag{10}$$

where $A \in \Re^1_{++}$ is such that

$$A < T \in \Re^1_{++}$$

The models needed for stipulating the historical operators $F_i(X, U, t)$ may vary from the truly simple to the very detailed. These operators need not be smooth, and can in fact be based on numerical results of separate simulations or time series analyses of real data.

5.2 Equivalent Extremal Problems and Line Integrals

For a variational inequality of the type (1), an equivalent extremal problem is

$$\min Z(u) = \sum_{i=1}^{m} \int_0^T \oint_0^u F(y, u, t)\, dy dt \tag{11}$$

$$u \in U \in (L[0, T])^m \tag{12}$$

where the Hilbert space $\mathcal{H}$ for this problem is $(L[0, T])^m$, y is a dummy variable of integration, $F(\cdot, \cdot, \cdot)$, $f(\cdot, \cdot, \cdot)$ and $\Gamma(\cdot, \cdot, \cdot)$ are convex in (x, u) and weakly continuous on U, and the line integral $\oint_0^u$ must of course be well defined.[7] The line integral (11) is familiar from static finite dimensional equilibrium modeling where it is extremely problematic since it is not generally single valued and has a value dependent on the path of integration. However, in the present context, the dynamics within U lead to

$$dy = f(y, u, t)\, dt \tag{13}$$

A simple substitution of (13) into (11) leads directly to

$$\min Z(u) = \sum_{i=1}^{m} \int_0^T \left[\int_0^t F(x, y, \xi)\, f(y, u, \xi)\, d\xi \right] dt \tag{14}$$

where is ξ is another dummy variable of integration. Evidently (14) is an ordinary integral. Consequently, any DVI formulation of noncooperative dynamic network equilibrium can be solved directly by finite dimensional mathematical programming if a finite difference approximation is employed or by feasible direction and projection methods in function space if continuous time is employed since the presence of line integrals poses no particular difficulty.

5.3 Stackelberg and Other Moving Equilibria

One may also extend (1) by postulating alternative forms of competition among the game agents (other than the CNB noncooperative assumption). Of course one possibility is the well-known leader-follower framework of Stackelberg-Cournot-Nash-Bertrand (SCNB) games that postulates a single omniscient agent (the leader) and describes the remaining agents as CNB

---

[7] $(L[0, T])^m$ is the m-fold product of the space of square integrable functions for $t \in [0, T]$. We also implicitly assume a regularity condition adequate to make the Kuhn-Tucker conditions in the relevant Hilbert space valid.

players. The Stackelberg assumption leads to a hierarchical optimal control problem,[8] with the outer problem being a classical optimal control problem and the inner problem a variational inequality like (1). However, there is no theoretical reason that several additional vertical and horizontal hierarchies could not be modeled. In fact Thornton (1995) applies this sort of reasoning to develop a multilevel dynamic model of US Army personnel management and promotions.

To model games that are not purely (non)cooperative, we may add to the constraints of U the constraints defining the core of the game. This will result in further constraining U so that

1. individual and group rationality hold; and
2. Pareto optimality for the grand coalition holds.

The core of the game is the most general notion of solution of an arbitrary game; most real-world games may be viewed as enforcing the core constraints as well as constraints peculiar to the specific institutional framework being studied. As Harker (1985) points out, enforcing the core makes the resulting problem min-max in nature. For finite dimensional problems it has been possible to develop satisfactory computational schemes for core-constrained general games through the use of relaxation and duality, so that the subproblems encountered are conventional VIs. There is every reason to expect that similar computational schemes could be developed for general games in function spaces for which the subproblems will be DVIs.

Still other types of non-traditional hierarchical formulations of (1) extending the Stackelberg notion of leaders and followers can be created. In particular, one could allow sets of Cournot-Nash-Bertrand agents to be hierarchically situated relative to other sets of CNB players. Also, empirically derived or conjectural response functions can be embedded that allow game behaviors among agents that have no classical analogues. This is because the structure of (1)–(7) allows any “rule” to be embedded that can be stated in terms of the state and control variables.

The mathematical formulations resulting from the Stackelberg and *ad hoc* reaction function perspectives we have described above are nonlinear variational inequality constrained mathematical programs in either $\Re^n$, if a finite difference approximation of the DVI is employed, or in Hilbert space. As such, well known classical numerical methods can be applied to adduce local solutions (the problems will be nonconvex). Alternatively, more modern meta heuristics, AI and genetic algorithms can be employed. See in particular Friesz et al. (1992), Friesz et al. (1993), Friesz and Shah (1999), Friesz and

[8] We omit the symbolic statement of the dynamic SNCB model for the sake of brevity. See in particular Friesz and Shah (2001b).

Shah (2001a) and Friesz and Shah (2001b) for a discussion the application of nontraditional algorithms, especially AI, to variational inequality constrained mathematical programs occurring in the context of transportation networks. See Luo, Pang and Ralph (1996) for a discussion of classical methods for such problems.

5.4 Extension to a Stochastic Setting

The ability imparted by (10) to plan based on future information may be stochasticized to account for intrinsic uncertainty, although we avoid for this overview paper the mathematical detail involving the Itô calculus that such an extension necessarily entails. Instead we merely comment that substantial success has been reported in the financial engineering literature regarding the numerical solution of infinite dimensional variational inequalities through finite difference approximations. The stochasticized DVI models we envision will be similar in mathematical structure to the option pricing stochasticized VIPs; so, it is reasonable to expect that we can obtain similar numerical success with finite difference methods for stochasticized DVIs. There is also reason to be hopeful that direct use of feasible direction and projection methods in Hilbert space will be effective; see in particular the excellent summary of mathematical programming and variational inequality algorithms in Hilbert space contained in Minoux (1986) as well as Luenberger (1969).

6. Disequilibrium Modeling

In the previous section, we considered identical agents and modeled a time varying or so-called *moving equilibrium*. However, there is no *a priori* reason that the DVI must be constrained by dynamics (8) that presume a moving equilibrium obtains. One may, instead, postulate adjustment processes that correspond to *equilibrium tending* behaviors. Dynamics based on such adjustment processes are called *disequilibrium dynamics* and allow equilibrium to be approached only in the limit of time $T \to \infty$; furthermore, such models may also be formulated in terms of relaxed notions of equilibrium/stability in the sense of Lagrange.[9] Because dynamics of this sort recognize that there may be additional constraints (beyond the dynamics themselves) that need to be enforced for the disequilibrium trajectories to be realistic, provision must be made for the inclusion of technological, doctrinal, and policy constraints

[9] By *stability in the sense of Lagrange* we mean that the trajectories stay within a bounded region of state space after the passage of enough time, but convergence to a stationary point or periodic attractor does not necessarily occur. This notion of stability admits chaotic attractors as generalized equilibria.

that modify the trajectories of states. The manner in which such constraints are treated distinguishes one disequilibrium dynamic model from another, and in fact we may identify two main schools of thought regarding the specification of disequilibrium dynamics:

1. classical non-tatonnement dynamics; and
2. projective dynamics.

The next two subsections briefly describe these two points of view.

6.1 Non-Tatonnement Dynamics

In order to understand the notion of *non-tatonnement dynamics* [see Varian (1984, pages 247–249) for a formal definition], we need to first clarify what is meant by *tatonnement dynamics.* Generally tatonnement dynamics are based on the assumption that the rate of change of each state variable is directly proportional to some measure of how far the current system state is from equilibrium. For example, the rate of change of price is proportional to the excess demand, where by excess demand we mean the amount by which demand exceeds supply. Since this difference may be positive or negative, price rises or falls in accord with the sign of excess demand. Although the name is perhaps poorly chosen, such tatonnement dynamics are referred to as *non-tatonnement dynamics* when steps are taken to make the disequilibrium states visited, prior to equilibrium, *correspond to actual observable states* of a system of interest. As Varian (1984) points out, this is done by making the state variables correspond to flows of *generalized commodities*[10] and introducing side constraints that ensure inventories and backorders are appropriately accounted for and modeled. Thus, classical non-tatonnement disequilibrium dynamic models take the form of systems of differential, difference or combined differential-difference equations with side constraints, and as such are naturally suited for appending to a pure optimal control problem or to a DVI.

6.2 Projective Dynamics

Projective dynamics are concerned with formally embedding the constraints of interest into a tatonnement model using the minimum norm projection for the relevant Hilbert space. See Smith et al. (1997) for a review and categorization of different types of projective dynamics. Although projective dynamics also use an "excess" of some state variable to model the disequilibrium states, they do not involve explicit side constraints beyond the dynamics themselves. Rather, the constraints pertinent to the system's description are used to define

[10] A *generalized commodity* is any flow type that is infinitely divisible or nearly so for the time-scale considered, as for example sugar, money, automobile trips, munitions and the like.

a manifold onto which the excess signal is projected to create differential equations whose right sides are nondifferentiable. It can be shown that such projective dynamics may be re-stated as differential inclusions and the growing body of functional analysis results for differential inclusions may be used to perform qualitative (stability, existence and uniqueness) analyses of such dynamic systems; see in particular the in-depth treatment of these issues by Aubin and Cellina (1984).

6.3 Building Disequilibrium Game Theoretic Models

Non-tatonnement dynamics and projective dynamics may be used to create disequilibrium dynamic extensions of the classical static Cournot-Nash-Bertrand (CNB) and Stackelberg-Cournot-Nash-Bertrand (SCNB) non-cooperative games that employ differential variational inequalities (DVIs). The key result regarding disequilibrium modeling with non-tatonnement dynamics is that a dynamic CNB equilibrium-tending game theoretic disequilibrium among non-cooperative agents takes the form of a differential variational inequality. If the dynamics are instead based on the notion of projective dynamics, we obtain a rather more challenging version of the DVI. In particular, when we generalize the notion of CNB equilibrium using projective dynamics, we obtain dynamics whose right hand sides are nonsmooth and which may be represented as differential inclusions [Aubin and Cellina (1984)]. If we similarly generalize the SCNB equilibrium, we obtain nonsmooth optimal control problems (or, equivalently, optimal control problems involving differential inclusions).

In addressing dynamic disequilibrium competition, we seek to solve dynamic CNB games stated as DVIs; the key relevant result, given in Friesz et al. (2001a), is:

THEOREM 4. *A dynamic CNB equilibrium-tending game-theoretic disequilibrium among agents described by the vector-valued disutility function $F(u) \in \mathcal{H}$ is equivalent to the following differential variational inequality: find $u^* \in U \subseteq \mathcal{H}$ such that*

$$\left\langle F\left(x^*, u^*, t\right), \left(u - u^*\right)\right\rangle = \int_0^T F\left(x^*, u^*, t\right)\left(u - u^*\right) dt \geqslant 0 \quad \forall u \in U \subseteq \mathcal{H} \tag{15}$$

where is a $\mathcal{H}$ Hilbert space and

$$\langle \cdot, \cdot \rangle \text{ denotes an inner product in } \mathcal{H} \tag{16}$$

$$u \doteq \left(u_i(t)\right)_{i=1}^m \tag{17}$$

$$U = \left\{ u : \frac{dx}{dt} \in f(x, u, t), \Gamma(u, x, t) = 0, x(0) = x^0 \right\} \tag{18}$$

$$F : \mathcal{V} \times U \times \Re^1 \longrightarrow \mathcal{H} \tag{19}$$

$$f \in \mathcal{V} \times U \times \Re^1 \longrightarrow \mathcal{H} \tag{20}$$

$$\Gamma : \mathcal{V} \times U \times \Re^1 \longrightarrow \mathcal{H} \tag{21}$$

for $F(\cdot, \cdot, \cdot)$, $f(\cdot, \cdot, \cdot)$ *and* $\Gamma(\cdot, \cdot, \cdot)$ *convex in* (x, u) *and weakly continuous on* U, *and* $x(u)$ *a mapping from* U *to* $\mathcal{V}$, *also a Hilbert space.*

Note that $f(x, u, t)$ is now a set and $f(x, u, t) \in 0$ is an inclusion that describes the equilibrium obtained in the limit $T \to \infty$.

SCNB models of disequilibrium employ DVIs like (15), together with the relevant constraints cited above to model the CNB agents and an extremal problem to model the Stackelberg agent. The result is a bilevel model in the form of an optimal control problem with DVI constraints. This is the most difficult class of DVI disequilibrium models from the point of view of analysis and computation. Not to be confused with bilevel static SCNB games, these infinite dimensional bilevel models are essentially unstudied as of the present. Again the constraints defining the core of a general game may be appended to U in order to describe other than pure noncooperation or pure collusion in disequilibrium games.

7. Candidate Case Studies in Network Science

The modeling and computational approaches we have discussed above lend themselves to two very timely case studies, namely

1. **Economic Planning and Capital Budgeting in the Context of Limited Warfare**. This important application is discussed in some detail in the next section, but in a nutshell may be described as planning for warfare that does not destroy the civilian economy in the theater of operations
2. **Internet Vulnerability**. The Internet may be viewed as a competitive dynamic system for which agents route individual packets of data or voice over a network and possible links that these packets employ. A behaviorally correct and computationally tractable mathematical model of the Internet will allow the study of network vulnerability with an aim of neutralizing the effects of sabotage and economic warfare.

Each of these applications necessitates a dynamic model of competition, as we describe below.

7.1 Economic Planning and Capital Budgeting for Networks in the Context of Limited Warfare

It is now a recognized military doctrine that the United States and its NATO allies should be prepared to fight simultaneously in multiple limited engagement regional wars. A hallmark of this thinking is that collateral damage to the civilian economy and its underlying infrastructure should be limited by previously stipulated bounds, that will of course vary from conflict to conflict. The desire to limit the deleterious impacts of warfare on the civilian economy stems from the recognition that the post-war health of the economy where the conflict occurs has a substantial impact on the total length of U.S. deployments and subsequent economic aid. This doctrine of limited warfare means that the problems of targeting infrastructure for disruption or destruction is intrinsically multi-objective, hierarchical and game theoretic.

Because of the nonlinearities arising from infrastructure congestion externalities, one cannot generally know by inspection or back-of-the-envelope calculations, the full impact of degrading a given infrastructure component. This nonlinearity of infrastructure is well recognized in network planning and optimization, where it is referred to as the *Braess Paradox* (BP). Briefly stated the BP concerns the possibility that increasing the capacity of a given infrastructure component may lead to an unanticipated network wide increase in congestion. For the limited warfare application, BP is stated in a contrapositive fashion as: degrading a specific infrastructure component may lead to localized increases in efficiency that prevent an overall degradation of efficiency; indeed it is even possible that efficiency will be globally enhanced for certain periods of time. This seeming paradox arises from the potential for network game agents to redistribute activities from the damaged components to decentralized locations that actually provide better service coverage.

To answer the call for a means of predicting the impacts of infrastructure interdiction, one is tempted to say that some kind of traditional economic forecasting may be used to predict the health of the war-impacted civilian economy. Yet this is not so, primarily because economic forecasting is predicated on analyses that describe statistical trends, not the shocks associated with war-based destruction. Furthermore, large scale event driven simulations of the economy of a particular region are both expensive and possibly not practical if the region has not previously been studied in-depth. In fact, history is rife with examples of failures to properly anticipate the economic and behavioral consequences of warfare. Moreover, direct economic and information warfare without firing a shot may well be seen in the 21st century. Consequently, extended game theoretic models of the DVI type that simultaneously analyze military operations and the functioning of the impacted civilian economy are critical to the conduct of limited, economic and information warfare.

In fact the limited warfare problem can be viewed as a special kind of economic planning/capital budgeting problem of the DVI Stackelberg-Cournot-Nash-Bertrand type wherein certain infrastructure is identified for surgical removal. This "disinvestment" is constrained by military considerations as well as the need to preserve some level of functioning of the civilian economy. The limited warfare problem is intrinsically a disequilibrium problem since the duration of transient effects is of great importance. Moreover the limited warfare problem is also a hierarchical one in that, during the early phases of a campaign, military objectives will clearly predominate, while the return to economic equilibrium and the level of equilibrium prices will be of key concern after force withdrawal.

7.2 Internet Vulnerability

The Internet is already a central feature of economic and intellectual life in the United States and seems destined to assume a similar role around the globe. Consequently, the vulnerability of the Internet and of component local area networks connected to it to failures and attacks is an issue of increasing importance to national security. Nonetheless very little effort has been expended to date to develop strategic plans for the evolution and protection of the Internet. This circumstance is largely due to the distributed nature of data and services spawned by the Internet, whereby no single entity can be identified as responsible for taking such a strategic view. In the language we have introduced above, the Internet is a dynamic, many-agent mathematical game wherein certain Stackelberg-like agents who control critical hardware and protocols have great power, while others are atomistic Cournot-Nash-Bertrand users with little power.

A dynamic game-theoretic model of the Internet is needed in order to correctly account for transient effects that may be of great policy concern, such as attacks on large commercial Internet service providers that disrupt the national economy, as well as attacks that seek to cripple specific military computers. Furthermore, the Internet does not have conventional static equilibrium states, but instead must be modeled using the Itô calculus, stochastic differential equations and variational inequalities that do not presume stationary states but allow self-similarity and so-called moving equilibria. Numerical solution of the resulting dynamic model will be necessary, as there is essentially no prospect that it can be solved in a closed form. Moreover, optimization models pertinent to Internet design and control will necessarily be constrained by these dynamics; as such Internet optimization models will likely require algorithms that are a hybridization of simulation, heuristics and numerical analysis techniques perhaps in a way not used before.

Critical to the construction of any model addressing the vulnerability of the Internet is the need for a dynamic description of packet flows/packet switching over the Internet. Such dynamics must necessarily reflect the fact that, in the current Internet, the rate at which a source sends packets is controlled by the Transmission Control Protocol (TCP), implemented by software on the computers that are the source and destination of the data. The basic idea behind TCP is that when the network becomes overloaded, one or more packets are lost and the loss of a packet is taken as the indicator of congestion. The loss of a packet causes the destination to inform the source and the source slows its rate of transmission; TCP then orders a gradual increase in the transmission rate until congestion is again indicated.

Mathematical modeling of the dynamic process we have described and of the related decisions that lead Internet users to contact sources and initiate transmissions in the first place is by no means straightforward. Even so, substantial agreement is beginning to emerge regarding the critical features of the Internet that must be captured by an effective model. Existing literature [see, for example, Kelly (1997), Willinger and Paxson (1998), Gibbens and Kelly (1999), and Kelly (2000)] allows us to make the following observations about the Internet:

1. Like the stock market, there is an amazingly detailed and accurate, temporally and spatially comprehensive, data record of Internet usage at all levels of disaggregation, owing to the way TCP is implemented.
2. So-called burstiness describes Internet traffic whereby the natural time scale for which Internet transmissions display periods-of-greater than average activity and less-than-average activity is around 10 msec (in contrast with classical telephony for which the time scale is 100 msec). However, bursts of longer duration do occur on the Internet; indeed Internet traffic seems to involve multiple time scales.
3. Like the stock market, empirical evidence suggests that so-called power law distributions with high variability and fat tails describe the usage of individual sources and classes of sources on the Internet. That is to say, Poisson-type queueing models used in traditional mathematical telephony are not relevant. Indeed, empirical evidence indicates that the proper distributions must correspond to what some have called fractal-Gaussian noise. This self-similarity has been widely observed for Internet traffic, in that finer and finer time scale resolution yields qualitatively similar distributions so that a fractal nature is displayed. It is also likely that the fat tails of empirical distributions arise from long range (in time) dependence associated with substantial autocorrelations over time periods that are several multiples of the intrinsic time scale.
4. Data indicate that with near statistical certainty, aggregate Internet traffic grows exponentially. Although, Poisson-type models must be rejected as a foundation for describing the arrivals of individual data packets within

the Internet, there is very good evidence that Poisson models are quite appropriate for characterizing arrivals of humans to the Internet.

5. Statistical inference is not a reliable means of modeling the Internet, aside from forecasting an overall growth in its usage.
6. Simple deterministic differential equations will never be adequate for describing Internet dynamics.
7. Stochastic differential equations based on Brownian motion (Wiener processes) that are so popular right now in financial engineering are also not fully appropriate since they presume Gaussian distributions without fat tails [Mantegna and Stanley (2000)]. So one cannot without modification apply the main results of the Itô calculus.
8. Any equilibria realized on the Internet will correspond to very short time scales and, consequently, notions of moving equilibria and disequilibria must be included in any dynamic models of the Internet.
9. Optimization and game-theoretic models needed for Internet design and control will be dynamic, stochastic and require new algorithms and numerical methods.

A consequence of this emerging consensus is that "simple" mathematical models of the Internet are almost certainly doomed to have little explanatory power.

Rather it will be necessary in modeling the Internet to deal directly with its nonlinearity, stochasticity and dynamic nature that make it much more akin to the stock market than to classical telephony. Even so, it will be necessary to modify the stochastic differential equations/infinite dimensional variational inequality approach developed for modeling securities[11] before it may be used to model the Internet. In so doing, one is lead directly to formulations that view Internet users as noncooperative game agents whose realized rates of data transmission are state variables devolving from dynamics that view protocols as controls; these dynamics involve an intrinsic stochastic term that corresponds to Levy-stable distributions with fat tails; that is, one is led directly to stochastic DVIs.

8. Concluding Remarks

We have introduced the reader to the differential variational inequality framework and discussed its usefulness in modeling dynamic games. An introduction has been given to dynamic game theory and the different forms

[11] See in particular Wilmott et. al (1993) for an excellent introduction to how stochastic differential equations and infinite dimensional variational inequalities are applied to the study of options.

that these games may take. Furthermore, two case studies for ongoing and future research into limited warfare and the Internet were introduced which also take the form of dynamic noncooperative games and as such may be modeled using the DVI framework.

References

G. Anandalingam and T.L. Friesz (1992), Hierarchical Optimization: An Introduction, *Annals of Operations Research*, 34, pp. 1–11.

G. Anandalingam and T. L. Friesz (eds) (1992), *Hierarchical Mathematical Programming*, Baltzer AG Scientific Publishing Co.

J. P. Aubin and A. Cellina (1984), *Differential Inclusions*, Springer-Verlag, Berlin, 342 pp., 1984.

C. Baiocchi and A. Capelo (1984), *Variational and Quasivariational Inequalities: Applications to Free-Boundary Problems*, Wiley.

T. Basar and G. J. Olsder (1998), *Dynamic Noncooperative Game Theory*, 2nd Edition, SIAM, Philadelphia.

D. Bernstein and S. A. Gabriel (1997), "Solving the Nonadditive Traffic Equilibrium Problem," in *Lecture Notes in Economics and Mathematical Systems: Network Optimization*, P. M. Pardalos, D. W. Hearn, W. W. Hager, eds., Springer, 72–102.

M. Ferris and J. S. Pang (1997), Engineering and economic applications of complementarity problems, *SIAM Review*, 39, 669–713.

A. Friedman (1982), *Variational Principles and Free Boundary Problems*, Krieger.

T. L. Friesz and S. Shah (2001a), An Overview of Nontraditional Formulations of Static and Dynamic Equilibrium Network Design, *Transportation Research, 35(1).*

T. L. Friesz and S. Shah (2001b), Foundations of a Theory of Disequilibrium Design, paper 10 in *Transportation Networks: Recent Methodological Advances* (M. Bell, Editor), Elsevier Science Publishers, pp. 143–162, 1999.

T. L. Friesz, S. Shah, D. Bernstein (1999), Disequilibrium Network Design: A New Paradigm for Transportation Planning and Control, paper 6 in *Network Infrastructure and the Urban Environment* (L. Lundquist et. al, editors), Springer-Verlag, pp. 99–111.

T. L. Friesz, Z. Suo, D. Bernstein and R. Tobin (2001), Differential Variational Inequalities and Dynamic Network Equilibrium, *Networks and Spatial Economics*, 1(3).

T. L. Friesz et al. (1992), Simulated Annealing Methods for Network Design Problems with Variational Inequality Constraints, *Transportation Science*, 26(1), pp. 18–26.

T. L.Friesz et al. (1993), The Multiobjective Equilibrium Network Design Problem Revisited: A Simulated Annealing Approach, *European Journal of Operational Research*, 65, pp. 44–57.

S. A. Gabriel (1998), An NE/SQP Method for the Bounded Nonlinear Complementarity Problem, *Journal for Optimization Theory and Applications*, 97, 493–506.

S. A. Gabriel (1998), A Hybrid Smoothing Method for the Mixed Nonlinear Complementarity Problem, *Computational Optimization and Applications*, 9, 153–173.

S. A. Gabriel and D. Bernstein (1997), The Traffic Equilibriuim Problem with Nonadditive Costs, *Transportation Science*, 31, 337–348.

S. A. Gabriel and D. Bernstein (2000), Nonadditive Shortest Paths: Subproblems in Multi-Agent Competitive Network Models, *Computational and Mathematical Organization Theory*, 6, 29–45.

S. A. Gabriel and J. S. Pang (1992), An Inexact NE/SQP Method for Solving the Nonlinear Complementarity Problem, *Computational Optimization and Applications*, 1, 67–91.

S. A. Gabriel and J. S. Pang (1994), A Trust Region Method for Constrained Nonsmooth Equations, in *Large-Scale Optimization: State of the Art*, edited by W.W. Hager, D.W. Hearn, and P. Pardalos, Kluwer Academic Publishers, 159–186.

S. A. Gabriel and J. J. More (1997), Smoothing of Mixed Complementarity Problems, in *Proceedings of the International Conference on Complementarity Problems*, The Johns Hopkins University, Baltimore, Maryland, 105–116.

S. A. Gabriel, A. S. Kydes, P. Whitman (2001), The National Energy Modeling System: A Large-Scale Energy-Economic Equilibrium Model, Theory and Computational Results, Vol. 29, *Operations Research*, 14–25.

R. J. Gibbens and Kelly, Resource Pricing and the Evolution of Congestion Control, *Automatica*, 35, 1999

P. T. Harker (1985), Investigating the Use of the Core as a Solution Concept in Spatial Price Equilibrium Games, *Spatial Price Equilibrium: Advances in Theory, Computation and Application*, Springer-Verlag (P. T. Harker, editor), 41–72.

P. T. Harker and J. S. Pang (1990), Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications, *Mathematical Programming*, 48, 161–220.

R. Isaacs (1965), *Differential Games*, First Edition, Wiley.

F. P. Kelly (1997) Charging and Rate Control for Elastic Traffic, *European Transactions on Telecommunications*, 8, pp. 33–37.

F. P. Kelly (2000), Mathematical Modelling of the Internet, working paper, Statistical Laboratory, University of Cambridge.

D. Kinderlehrer and G. Stampacchia (1980), *An Introduction to Variational Inequalities and Their Applications*, Academic Press.

V. K. Le and K. Schmitt (1997), Global Bifurcation in Variational Inequalities: Applications to Obstacle and Unilateral Problems, Springer Verlag.

Z. Q. Luo, AJ. S. Pang and D. Ralph (1996), *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press.

D. G. Luenberger, *Optimization by Vector Space Methods* (1969), Wiley.

R. Mantegna and H. Stanley (2000), *An Introduction to Econophysics: Correlations and Complexity in Finance*, Cambridge University Press.

M. Minoux, Mathematical Programming: Theory and Algorithms (1986), Wiley.

J. S. Pang and S. A. Gabriel (1993), NE/SQP: a Robust Algorithm for the Nonlinear Complementarity Problem, *Mathematical Programming*, 60, 295–337.

P. A. Samuelson (1947), *Foundations of Economic Analysis*, Harvard University Press

T. E. Smith, T. L. Friesz et al. (1997), A Comparison of Two Minimum Norm Projective Dynamic Systems and Their Relationship to Variational Inequalities, *Complementarity and Variational Problems* (M. Ferris and J.-S. Pang, editors), SIAM, pp. 405–424.

P. Thornton (1995), *A Hierarchical Mathematical Programming Model of the U.S. Army Promotion and Assignment Process*, Ph.D. dissertation, George Mason University.

H. R. Varian (1984), *Microeconomic Analysis*, 2nd Edition, Norton.

W. Willinger and V. Paxson (1998), Where Mathematics Meets the Internet, *Notices of the American Mathematical Society*, 45, pp. 961–970 (www.ams.org/notices/).

P. Wilmott, J. Dewynne and S. Howison (1993), *Option Pricing: Mathematical Models and Computation*, Oxford Financial Press.

Chapter 5

Characterization and Monitoring of Nonlinear Dynamics and Chaos in Manufacturing Enterprise Systems

S.R.T. Kumara[1] and S.T.S. Bukkapatnam[2]

[1]Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA, USA

[2]Industrial Engineering and Management, Oklahoma State University, Stillwater, OK, USA

Abstract Much of the complexity in modern enterprises emerges from the nonlinear and likely chaotic dynamics of the underlying processes. These processes are defined over multiple scales of system granularity, for e.g., supply chain-level, through shop floors, down to a machine or a core physical operation level. Characterization of this complexity is imperative for improving predictability of quality and performance in modern physical and engineered systems. In this paper we present some theoretical developments and tools aimed at advancing the applications of nonlinear dynamic systems principles in manufacturing processes and systems, with specific emphasis on characterizing and harnessing chaos in these complex systems. We examine the current developments in addressing predictability in two important facets of a manufacturing enterprise, namely, process level characterization and monitoring, and systems level characterization. For each case, we concisely evaluate the relevant alternative approaches and layout certain open issues. We hope that this paper will spur further development of methodologies adapting nonlinear dynamics and chaos principles for advancing various aspects of manufacturing enterprises.

Keywords: intelligent agents; manufacturing; fractals

1. Introduction

A vast majority of real-world systems and processes are nonlinear. They exhibit a variety of behaviors depending on how these processes are initiated, and their emergent behaviors can be completely altered by introducing various forms of stimuli. Also, most physical systems exhibit aperiodic, strange, and irregular behaviors; periodic behavior is a rarity. However, for a long time, many engineering applications have been using linear models to analyze and control the underlying processes. This is because, accurate nonlinear models are perceived to be computationally intractable. Many theories of real-world systems including those in manufacturing are formulated based on a linear perspective.

However, linear models do not capture several critical and strange behaviors encountered in the real processes. With the advancements in computing power, researchers have now realized the importance of studying nonlinear dynamics and chaos associated with these processes. Concepts commonly bundled as chaos theory have facilitated the study of these "strange" and complex behaviors.

During the last decade, several investigations into characterization, monitoring and control of manufacturing enterprise systems, from the enterprise-level down to a machine level, have started to take a nonlinear viewpoint. This paper reviews some of the recent advancements in the application of nonlinear dynamics in manufacturing. Specifically, we plan to present some of the authors' decade-long investigations into characterization, monitoring, and control of manufacturing systems at the process level (Section 3), as well as characterization and performance prediction of a complex manufacturing system (Section 4). We hope that our limited and focused review will exemplify the advancements being made in the application of nonlinear dynamics to several other manufacturing enterprise processes and aspects therein.

2. A Primer to Relevant Aspects of Nonlinear Dynamics

Many physical systems, including manufacturing enterprises that produce continuous-time response, may be modeled by a set of differential equations of the form

$$dx/dt = F(x)dt \tag{1}$$

where $F(x)$ is generally a nonlinear *vector field*. The solution to (1) results in a trajectory

$$x = f(x(0), t) \tag{2}$$

where f represents the flow that determines the evolution of $x(t)$ for a particular initial condition $x(0)$.

If the system is dissipative, as the system trajectories evolve from different initial conditions, the solutions usually shrink asymptotically to a compact (closed and bounded) subset of the whole state space. This compact subset, representing the steady-state of the system, is called an attracting set. Every attracting disjoint subset of an attracting set is called an *attractor* (Nayfeh, 1995; Moon, 1987). For specified operating conditions, represented by a process parameter vector p, the dynamics may have multiple attractors. Unlike linear systems, the different attractors may portray significantly different types of behaviors ranging from static, quasiperiodic to chaotic.

As the attractors usually remain bounded, the flow exhibits a recurrent pattern, where trajectories starting from near-by points within an attractor may get separated exponentially as the system evolves. This condition is known as the sensitive dependence on initial condition (SDIC), and the attractor exhibiting SDIC is called a *chaotic* attractor. Trajectories from chaotic attractor are irregular and aperiodic with a continuous broadband spectrum. Chaotic attractors usually exhibit a special property of self-similarity or scale invariance – i.e., the response appears similar over multiple scales of observation. These scale invariant entities are commonly known as *fractals*.

An attractor may be locally stable or unstable depending on whether a trajectory initiating from an immediate vicinity of an attractor converges thereto, or diverges therefrom. Behavior of a nonlinear process may undergo an abrupt change as p is varied. This phenomenon is called a *bifurcation*. The value of p at which bifurcation occurs constitutes a bifurcation point. Bifurcation diagrams graphically delineate the influence of p on the dynamics, whose behavior is usually represented by a suitable "behavior functional" such as the amplitudes or periodicity of different orbits. If a bifurcation can be captured by means of local analysis around a simple attractor, it is said to be local; otherwise the bifurcation is said to be global. Local bifurcations are much easier to analyze. They occur whenever the linearized flow about an equilibrium point has exactly one purely imaginary eigenvalue. Global bifurcations take place in the presence of null eigenvalues and/or many purely imaginary eigenvalues. Global bifurcations result in very complex patterns. Their analysis is not tractable using currently available mathematical and scientific tools. However, state space reconstruction techniques (Nayfeh, 1995) may turn out to be extremely useful in capturing global bifurcations.

3. Process-Level Characterization, Monitoring and Control of a Manufacturing System

Machining is one of the most ubiquitous operations in a manufacturing system. About 70% of products produced in modern manufacturing enterprises are known to undergo machining process at some stage of their production. Outcome of this process oftentimes determines the quality and the operational performance of products emerging from a manufacturing enterprise system. Machining dynamics is the coupled dynamics of cutting process (CP) and machine tool structure (MTS). Variations in the uncut chip thickness resulting from vibrations, in turn, cause cutting force variations, thus sustaining machining dynamics (Tobias, 1965). Apart from dynamic cutting forces resulting from uncut chip thickness variations, disturbances due to chip breakage, chip interference, non-homogeneity of the work material, etc., contribute to vibrations in machining.

Local instabilities in machining dynamics may lead to an anomalous condition called chatter, which often manifests as an accentuation in vibration amplitudes, at times accompanied by a characteristic sound (Tobias, 1965; Koenigsberger, 1970). Chatter is the chief determinant of quality during precision machining of critical components for aerospace and medical applications. Chatter leads to poor surface finish, surface integrity and dimensional accuracy of the workpiece. It promotes cutting tool wear and other modes of cutting tool failure, damages the machine tool, and renders the work-area more prone to accidents. Chatter cannot be completely eliminated due to the presence of material and geometrical inhomogenities. Hence, modeling and control techniques are essential for improving performance.

Several researchers have studied the initiation and growth of chatter in machining through both purely experimental as well as lumped dynamic modeling (see Ehmann, 1997 for a recent review). For a long time, chatter was proposed to occur mainly due to the regenerative effects and mode coupling (Tobias, 1965; Koenigsberger, 1970; Albercht, 1965; Minis, 1990), assuming that machining dynamics is linear. Regenerative chatter can occur when machining is performed on a previously machined surface. The surface undulations left by the previous cut cause linear resonance conditions, thereby accentuating the cutting tool vibrations. Mode coupling chatter results from the resonance conditions prevailing due to the interaction between the cutting force components acting in different directions. These propositions have been able to explain many behavioral patterns of chatter, and they have led to the development of many monitoring and control paradigms.

Almost contemporaneously, nonlinearities in machining dynamics were recognized (Tlusty, 1981; Grabec, 1988; Lin, 1991). Tobias and his group modeled cutting forces as a nonlinear (polynomial) function of the delay term $(x_1(t) - x_1(t - \tau))$, and its higher order derivatives. This model has formed the basis for a few recent nonlinear dynamics-based studies of machining dynamics and chatter. Tlusty and Ismail (Tlusty, 1981) proposed that the finiteness of chatter amplitudes occurs due to momentary losses of workpiece−tool contact and concomitant interruptions of the driving force. These were among the earliest recognitions of the nonlinearities in machining dynamics. The so-called secondary nonlinear effects like shear angle variations, ploughing phenomena, rake face processes, and chip segmentation have also been recognized to be part of cutting process dynamics that can lead to chatter (Albercht, 1965; Komanduri, 1981).

Grabec (Grabec, 1988) proposed a two degrees-of-freedom nonlinear model to describe machining dynamics. He took coupled nonlinear variation of cutting forces relative to tool deflections and its higher order terms into account, but did not explicitly consider regenerative effects. Next, Lin and Weng (Lin, 1991) considered the variation of the shear angle due to the regenerative effect and improved Grabec's nonlinear model. Bukkapatnam et al. (Bukkapatnam, 1995) provided extensive experimental evidence that machining dynamics is low-dimensional chaotic under normal operating conditions. Nayfeh et al. (Nayfeh, 1997) studied Tobias' model, and characterized a subcritical hopf bifurcation as the source of chatter. They also demonstrated, through experimentation with boring process, that the post-bifurcation dynamics has multiple attractors which are consistent with a subcritical Hopf bifurcation (Nayfeh, 1995). Kalmar-Nagy et al. (Kalmar-Nagy, 1999) studied the same model and characterized the pre- and post-bifurcation attractors. They have also characterized certain nonlinear phenomena observed during machining operations.

However, most of the recent efforts concentrate on regeneration as the source of chatter. In reality, machining dynamics exhibits a rich set of dynamic behaviors. The current understandings and available models fail to explain many behavioral patterns of the vibration and force signals obtained from actual experiments (Bukkapatnam, 1999). These observations provide evidence for the existence of multiple bifurcations in dynamics and various similar manufacturing processes (Bukkapatnam, 2005). Future research in this direction must address the application of nonlinear dynamics to characterize and analyze these behaviors in order to improve predictability and monitoring of a process, and hence the system performance.

3.1 Process Monitoring

Monitoring requirements for nonlinear processes and systems are much more demanding than those for a linear process. For example, Fourier analysis for the most part may not be of avail. Process monitoring usually involves the following six subtasks:

1. ***Characterization:*** Assessing, through preliminary experimentation along with statistical and nonlinear dynamic hypothesis testing, the behaviors emerging as a consequence of complex system dynamics. The results of characterization determine the exact methodologies to be employed in the subsequent stages of monitoring.

2. ***Data acquisition***: Selecting appropriate sensors, designing and performing experiments, and collecting sensor data in a form amenable to further analysis.
3. ***Signal representation:*** Expressing/modeling a signal in terms of the components of certain basis vectors, so that understanding of patterns contained in the signal and the procedure for feature extraction will be much simpler and efficient in the transformed space.
4. ***Signal separation:*** Filtering out undesired signal components (contaminants) from the measured signal so that the accuracy of state estimation can be improved. This subtask is also called denoising.
5. ***Feature extraction:*** Extracting certain model parameters/coefficients from a properly represented signal. The extracted features should be sensitive to the state variables to be estimated.
6. ***State mapping:*** Associating the extracted features with an appropriate representation of unknown state variables.

A typical monitoring system may not involve all the above six subtasks. This is because the monitoring methodology, and the techniques employed therein vary significantly with the characteristics of the process dynamics and sensor signals. If the measured sensor signals are stationary, ARMA-type standard time-series models will be sufficient for signal representation. However, the measured signals from a manufacturing process or a system are hardly stationary because of the complex nature of the underlying dynamics. Furthermore, if the dynamics were to be linear, the response consists of finite multiple harmonics, and may be parsimoniously represented using Fourier representation. If the process dynamics is nonlinear, depending on the type of nonlinearity, nonlinear function approximations such as radial basis functions, wavelet transforms and neural networks may be employed for parsimonious signal representation. In addition, if the process dynamics exhibits low-dimensional chaos, the response is usually fractal-like. In such a case, parameters such as fractal dimensions serve as effective signal features. A list of these alternative techniques for performing different monitoring tasks is presented in Table 1. A detailed description of the individual subtasks is provided in the following subsections.

Process Characterization: The twin assumptions that the dynamics of a given process or system is chaotic and the sensor signals are not highly contaminated[1] underlie the application of fractal analysis. If these assumptions are not valid, the values of parameters/coefficients calculated using fractal analysis have little relevance to the characteristics of sensor signals. An experimental validation of the assumption of the dynamics as being chaotic

[1] Contamination refers to both measurements noise and dynamic contamination.

Table 1. Different techniques for process monitoring – A summary

Monitoring subtask	Technique	Assumptions/conditions for application
Data acquisition	FFT	Linear system with stationary additive noise
	DWT	Nonstationary, piecewise linear models
	Nyquist sampling	Linear system
	highest sampling rate	Nonlinear, usually chaotic, systems
Signal representation	Fourier analysis	Linear systems with additive noise, harmonic response
	Wavelet analysis	Nonstationary signal, piecewise linear system
	Karhunen-Loeve expansion	Piecewise stationary stochastic process, Linear system
	ARMA-type models	Stationary signal, linear system
Signal separation	Band-pass filtering	Linear system with stationary additive noise
	Linear opt-filtering algorithms	"Known" additive noise
	Wavelet-based denoising	Nonstationary, possibly nonlinear signal
	Neural network	Unknown noise characteristics, known desired behavior
	Shadowing	Chaotic signal with known noise distribution
	Neighborhood method	Known desired signal characteristics
Feature extraction	Fourier coefficients	Harmonic signal, linear system
	Wavelet coefficients	Nonharmonic, nonstationary signal
	STFT	Piecewise stationary linear system
	ARMA	Stationary signal, linear system
	RBF	Nonlinear system, response with additive noise
	Fractal dimensions	chaotic systems
	"specialized" model coefficients	System nonlinear characteristics understood
	Kalman filtering	Linear system, stationary additive noise
State estimation	EKF, EBKF	Weakly nonlinear systems
	Neural networks	Unknown relationships

may be done using a battery of tests involving statistical and nonlinear dynamics principles. A typical sequence of experiments for characterization may be found in (Bukkapatnam, 1995; Bukkapatnam, 1999; Bukkapatnam, 2000; Bukkapatnam, 2005) For characterizing machining process dynamics, we designed and conducted experiments, and performed graphical analysis of the measured signals. The graphical analysis suggested that the dynamics may be chaotic. In order to ascertain this result, we developed/used (i) surrogate data test, (ii) quasiperiodicity test, and (iii) Lyapunov exponents test. The results of these tests clearly established that turning process exhibits low-dimensional chaos in the normal operating conditions. The main implication of this result is that the dynamics is controllable using the principles of chaos theory.

Data acquisition: If the process exhibits low-dimensional chaos, periodogram and Fourier analysis of sensor signals are, to the most extent, meaningless. The data acquisition system should possess the highest possible sampling rate. This is especially true if one is interested in visualizing the structure of an attractor.

The sensors should be selected such that the signals therefrom contain definite pattern which can be mathematically related to unknown state variables. For example, force and vibration signals contain patterns which that can be related to the flank wear. Hence for flank wear estimation, we used force and vibration sensor signals (Kumara, 1994).

Signal representation: Signal representation is the key step in process monitoring because in many instances signal representation directly leads to signal compression, separation and feature extraction. Signal representation is usually carried out using linear time-series models such as ARMA, Fourier and wavelet models. It consists of approximating a given signal by a combination of certain basis functions. The model parameters lie in a transformed space, and these coefficients sometimes serve as features. In signal representation, parsimony is of utmost importance. That is the signal must be accurately represented using minimum number of basis functions. In many instances, signal representation is considered to be synonymous with signal compression.

Signal separation: Any signal obtained from on-line sensors is laden with noise, which has to be properly filtered out. If the dynamics underlying the signals were linear, Fourier analysis would have been adequate. But a nonlinear measured signal can be contaminated by measurement noise and/or dynamical noise (extraneous dynamics). Therefore, separation of contamination from the measured signal becomes very difficult (Schouten, 1994). Techniques such as shadowing, wavelet denoising, and neighborhood techniques may be used when the signal emanates from a low-dimensional chaotic attractor.

For example, we used wavelet denoising method to perform signal separation on force and vibration sensor signals obtained from experimentation. Wavelet denoising method (Johnstone, 1995) involves performing a discrete wavelet tranform (DWT) of the measured signal, obtaining the threshold from the wavelet transform coefficients, performing soft-thresholding of those coefficients and inverse transforming-back to the time-domain. In our case (Bukkapatnam, 2000) the use of wavelet denoising enhances the performance of feature extraction.

Feature extraction: If the signal is stationary or the process dynamics is linear, the coefficients of the basis functions themselves serve as signal features. However, especially when the process dynamics is chaotic, choice of basis function is never "correct". Therefore, other properties of signal have to be used. For example, for signal obtained from the turning process, we used fractal dimensions of the measured signal as the features. The fractal dimensions constitute a very parsimonious set of features that can be related to the unknown state.

State mapping: The essence of state mapping involves developing a model to relate the extracted signal features to the unknown state variables. The type of model can range from a simple regression model and a differential equation model to a neural network model. The choice of the model depends mainly on

- desired rate of generation of the state estimates,
- need for global modeling,
- understanding of the mathematical structure of the underlying signal-state relationship, and
- level of accuracy desired.

For many manufacturing processes including the basic turning process, the underling process mechanism is not completely understood. Hence neural network models are common. For example, for continuous estimation of tool wear in the turning process, two neural network architectures have been proposed by the first author of this paper and his group in the Intelligent Design and Diagnostics Research (now, LISQ) Laboratory, Pennsylvania State University.

The first method, developed by Kamarthi (Kamarthi, 1994) involved relating the signal feature with flank wear using a hybrid radial basis network. The signal features corresponded to the coefficients of ARMA model of force and vibration sensor signals and some coefficients in the wavelet representation of acoustic emission signals. Flank wear was represented by the height of the flank wear land.

The second method developed by Bukkapatnam et al. (Bukkapatnam, 2000) used a simple two-layer feed forward neural network to relate the signal features to the flank wear. The signal features consisted of the values of fractal dimensions computed from the measured signal.

3.2 Specific Applications

In this section, we will concisely present some research highlights on using nonlinear dynamics for sensing a particular type of nonlinear signal called Acoustic Emission (AE) occurring in machining process. AE refers to the high-frequency micro-elastic pulses generated in a material undergoing plastic deformation. In machining, deformation and/or friction at the shear zone, the rake and the flank faces are the three main sources of AE. Chip entanglement, crack formation and propagation, and chip breakage are also known to be AE sources (Byrne, 1995).

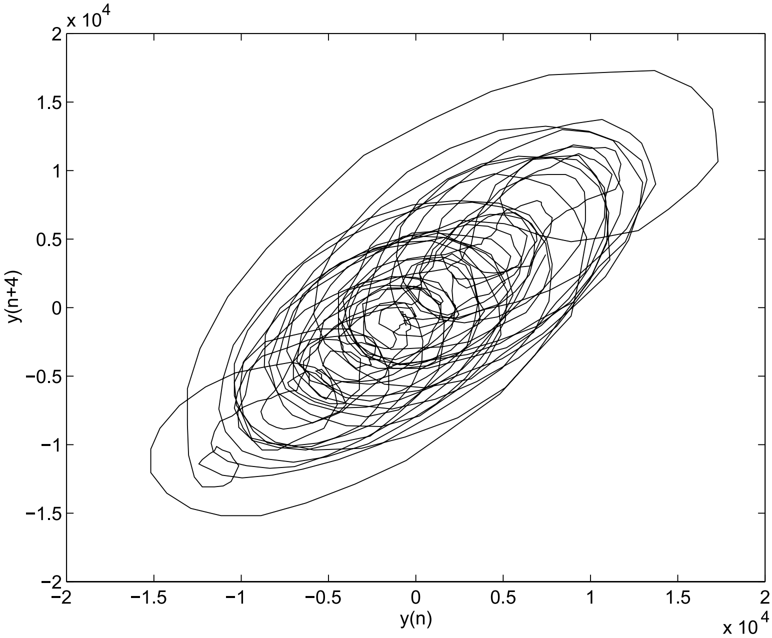


Figure 1. Recurring patterns in a representative lag plot of AE signals reveal a finite-dimensional attractor.

In the recent past, AE has emerged as a promising means to monitor machining processes (Byrne, 1995; Blum, 1990; Kannatey-Asibu, 1981; Liu, 1996), especially to detect breakage and other catastrophic failures. Recently, Bukkapatnam et al. analyzed AE signals collected from four different laboratories across the USA and Europe using nonlinear analysis methodologies (Bukkapatnam, 1999). They have found consistently recurring patterns underlying high frequency AE components, which were hither to categorized as noise. A sample lag plot of our findings is shown in Figure 1. Their characterizations have confirmed that AE signals exhibit chaotic behavior, and these high-frequency AE components emanate from a finite-dimensional attractor of a nonlinear process. This result implies that the common practice of filtering out high frequency non-stationary AE components as noise may not be appropriate.

Building on this result, Bukkapatnam et al. designed an extremely compact Suboptimal Wavelet Packet Representation (SWPR) for AE (Bukkapatnam, 1999), necessary for building real-time control models. This method was recently enhanced to yield customized basis functions for AE (Knapp, 1998), which was one of the first applications of lifting scheme in manufacturing.

Another challenge associated with sensing of nonlinear signals such as AE is noise suppression. Unlike linear signals, noise may occur in the same frequency bands as the desired signal, and hence band-pass filtering may not be appropriate. Bukkapatnam et al. developed a modified wavelet method (MWM) to address noise suppression in AE and similar signals. An MWM-separated signal strongly converges in mean to the best linear predictor of a noise-free trajectory. Also, it will remain bounded relative to the noise-free trajectory even in the presence of low multiplicative noise (Bukkapatnam, 2000).

The nonlinear quantifiers such as fractal dimensions extracted from MWM-separated signals serve as excellent features for on-line state estimation. Bukkapatnam et al. designed an over 90% accurate continuous estimator of flank wear in machining using these fractal quantifiers (Bukkapatnam, 2000), which capture the variation of machining dynamics with flank wear. Wavelet coefficients can be used to estimated energy distributed over various scales of resolution. These scale limited energies can serve as effective features to detect various faults in a system (Kumara, 1999)

In addition to monitoring, nonlinear dynamics principles can significantly advance process control. For example, chatter in machining, manifested as a limit cycle, was suppressed through simple feedback linearization (Kalmar-Nagy, 1999). Bukkapatnam et al. developed a robust control Lyapunov function (RCLF) using their characterization of nonlinearity in machining (Bukkapatnam, 1999). The control input synthesized based on this RCLF was found to suppress tool vibrations under both periodic and chaotic regimes, and was found to suppress chatter faster than the input synthesized using linear models.

Also, manufactured surfaces often exhibit fractal-like characteristics (Bhushan, 1995; Whitehouse, 1994). Fractal models provide insights on various functional and operational behaviors of surfaces well as surface-bearing components (Brown, 1998). Several quantifiers, including various fractal dimensions, structure functions, energy functions, scaling functions, as well as parameters such as power-law coefficients and lacunarity have employed to characterize these surfaces. Structure functions computed through multi-scale representations have also been found to be adequate quantifiers. The various fractal dimensions computed from profiles and/or surface patches are the most commonly used (Brown, 1993). These quantifiers have shown remarkable correlation with certain functional characteristics, specifically those of worn and fractured surfaces (Whitehouse, 1994). Apart from exhibiting multifractal properties, surface profiles have been speculated to be signatures of chaotic processes (Stark, 1999). The recent nonlinear analysis methods can further enhance surface characterization, modeling, and, to an extent, control.

4. Manufacturing Systems Characterization and Monitoring

Increasing global competition and eroding margins for manufacturers have been turning many products into commodities. More than ever, customers are demanding lower prices, higher quality, new and customized products. Manufacturing in such challenging environment requires new approaches for its design and operations. Because of the need for flexibility and fast adaptability of the production schedule to changes in demands, a more dynamic, rather than a static view of the production systems is needed. The advent of inexpensive sensor and data-logging technologies enable the manufacturing plants to operate in data-rich environments. These vast amounts of data can be harvested to build dynamic manufacturing models that offer the opportunities for realistic feedback and real-time control. This can help the system to respond in a fast and flexible manner to unplanned events in real-time.

The objective of this study reported in this section is to present an approach to model a manufacturing system that can capture the underlying nonlinear stochastic dynamics so one can gain a dynamic view of the system performance in terms of key variables such as work in process (WIP), backlog and throughput (Earni, March 2006). The resulting model can also help to track variations in the performance of a manufacturing system due to changing external and internal conditions. This in turn will help to design/redesign appropriate inventory and supply management policies. Also, by clearly understanding the dynamics of a manufacturing system, one would be able to run "what-if" scenarios to fully understand the implications of a disturbance (like machine break down, or rush order) in the system, and react to them more proactively instead of reactively. As an initial step in this direction, a simple discrete time sampled flow model, based on fundamental production control principles, is presented. It encapsulates a pull based manufacturing system. Based on this model, different control theoretic analyses were attempted to gain insights to better understand the performance of this system.

4.1 Manufacturing System Dynamics Modeling

The past research efforts on dynamic behavior of systems, specifically manufacturing systems, used six different approaches. Most of the research on improving the dynamic behavior of individual manufacturing enterprises is mostly focused on supply chain management and may be broadly categorized as follows.

Management games such as the Beer Game (Sterman, March 1989) are useful tools to illustrate the benefits of different supply chain strategies. The beer distribution game mimics the ordering and production decisions of a supply chain and let the players decide how much products need to be ordered to fulfill the demand. This game illustrates the effect of delays on order processing, production and shipping. These games provide anecdotal evidence and are a good learning device although they cannot be used for verifiable design of these systems. Other authors like van Ackere et al. (Ackere, 1993); Kaminsky and Simchi-Levi (Kaminsky, 1998) have extended or computerized this beer game approach.

Statistical approaches provide insights about the impact of demand properties such as standard deviation and correlation, and properties such as lead-times and information paths on inventory costs. Statistical methods are often used to quantify performance of real situations. These methods however, fail to show how to reduce or eliminate the detrimental dynamic effects, such as demand amplification. Physical insights into the causes and effects of system structure on performance are rarely obtained from these approaches.

Industrial dynamics, also known as system dynamics advocated by Forrester (Forrester, 1961), is a method of investigating the dynamical effects in large nonlinear systems without resorting to complicated mathematical control theory based models. Towill (Towill, 1994) studied the industrial dynamics models within the context of living and planned supply chains as utilized successfully by adopting a holistic approach in which the basic disciplines of industrial engineering and business process reengineering are integrated into a comprehensive methodology (Towill, 1993). Dejonckheere et al. (Dejonckheere, 2000) developed a control theoretic approach to measure and mitigates the variance amplification of orders within order-up-to policies and proved that these policies will always generate a bullwhip effect. Recent significant contributions in this area include from (Chen, 2000; Chen, 2000; Popplewell, 1987)

Simulation approaches alone suffer from being cumbersome, time consuming and only provide limited insight (Popplewell, 1987), but they do have the advantage of being able to model nonlinearities while avoiding complicated mathematics. Previous work using simulation is very prolific and includes (Forrester, 1961) and (Coyle, November/December 1982), who studied traditional supply chain structures (Cachon, 1997) and Waller, Johnson and Davis (1999) who studied vendor managed inventory (VMI).

Pritschow and Wiendahl (Pritschow, 1995) presented a dynamic application of control theory for production logistics and studied the effect of short term market changes and disturbances on the performance. Kilpatrick (Kilpatrick, 2003) developed a comprehensive framework for improving production efficiency using lean manufacturing principles. A model was created to analyze the dynamics of linear distribution systems, and showed how lean manufacturing represents an opportunity to sidestep many previously insurmountable difficulties that arise as a result of producing to fill inventory levels. The current efforts have addressed the characterization and modeling of manufacturing systems as nonlinear processes (Parunak, 1991; Wiendahl, 1999; Bongaerts, 1997).

Researchers have identified and characterized certain types of bifurcations in these systems (Parunak, 1991). Van Brussel and his research group developed the concept of holonic manufacturing to combine optimized schedules with robustness to disturbances, and have developed algorithms and software to implement this concept (Bongaerts, 1997).

Scholz-Reiter and Mueller (Scholz-Reiter, 2000) studied closed-loop production systems supported by neural networks. In addition to the developed buffer inventory control with neural networks, the suitability of neural networks for control of throughput time in production systems was also examined. The neural networks influence the release time of orders and the level of buffer inventory at the work systems, due to the difference in set value of throughput time. Conventional production control methods are normally based on static models, Wiendahl and Breithaupt (Wiendahl, 1999) developed a new general concept for dynamic modeling of job shop productions using control theory methods. Two controllers were developed: distributed backlog controller and a central WIP controller to track and improve the performance of a manufacturing system. Kim et al. (Kim, 2004) proposed a discrete dynamic model of a single workstation that can be used to design and analyze control algorithms for closed loop production system that can track the performance, especially in response to disturbances such as rush orders and periodic fluctuations in capacity, while ensuring that dynamic behavior remains favorable and robust. Methods of control engineering, such as transfer function and frequency response analysis, are used to make to study and improve the dynamic behavior of the system. Ratering et al. (Ratering, 2003) proposed integrated methods of control engineering with methods of production engineering to improve robustness and performance of production systems, while making dynamic analysis tractable and improving the understanding of complex behavior. Wiendahl and Westkamper (Wiendahl, 2002) used control theory to link external and internal requirements of a production system with manipulated variables of this system. Cho and Prabhu (Cho, September, 2002) presented a continuous feedback control approach for real-time scheduling of discrete events in distributed manufacturing applications and demonstrated highly nonlinear and discontinuous dynamic behavior, specifically, when the production demand in the manufacturing system exceeds the available resource capacity, then the control system "chatters" and exhibits sliding modes. Duffie and Falu (Duffie, 2002) developed a control-theoretic analysis with closed-loop control of backlog and WIP.

4.2 Modeling and Implementation Details

The system considered is a three station one product manufacturing model and is built on the concept of pull based manufacturing. The demand as seen by

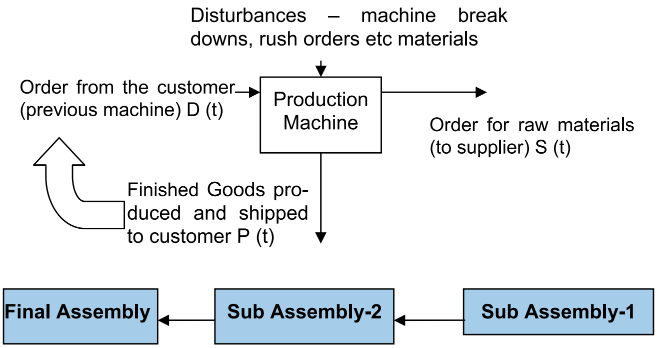


Figure 2. Schematic of a 3 stage assembly operation.

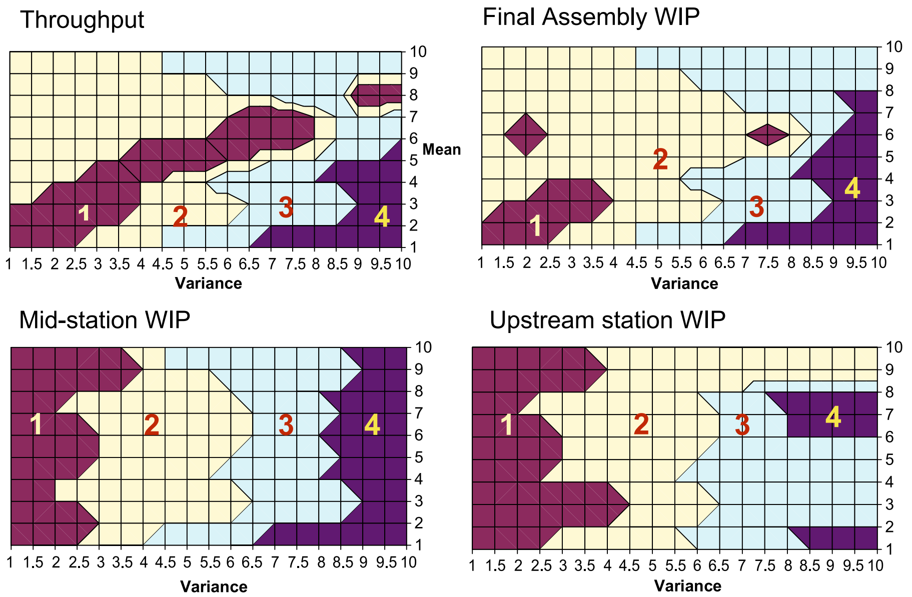
the final assembly operation (3) is translated into actual supplier requirements (Figure 2) and sent to the upstream supply requirements. Work-in-process (WIP), buffers before the workstation (jobs waiting for machine time), actual product delivered in response to the demand (throughput) and cycle time are some of the key performance indicators chosen to assess the performance of this manufacturing system.

The model uses two types of flows – information and material flow are modeled as discrete time functions. The information flow is the customer order information and signal authorizing the station to start production. The variable demand from the customer is received by the final workstation, and in response this workstation generates a Kanban signal to the downstream workstation to start production. This information is passed down the production sequence till the end of the line. In this model, we assumed that it takes one hour to relay the Kanban signal and two hours to actually deliver the products and an assembly time of one hour for each workstation. All times in the models are assumed to be deterministic. The initial conditions for all the variables is assumed to be zero, assumption was also made that all the capacities of the stations were known and equal for all times within the time horizon.

4.3 Results and Analysis

a. Behavior Analysis:

In this section, a comprehensive stability analysis of aforementioned manufacturing system is presented. Stability refers to the ability of a system return back to equilibrium following a perturbation (e.g. demand change). In other words, instability causes oscillatory (not necessarily periodic) behavior of increasing magnitudes over time. Unlike physical systems, in manufacturing systems global instability is rare since it is impossible for production rates to increase without bound. However, linear dynamical systems approaching



(a) (b)

Figure 3. Manufacturing system behavior plots for varying conditions of demand.

instability typically exhibit oscillatory behavior which, in this application, is manifested as demand amplification. So an appreciation of the configurations of operating conditions that lead to instability should enlighten practitioners to the mechanism of demand amplification (if the delay between the workstations is high) and lead to facilitate its attenuation. The performance of the manufacturing system is analyzed using the models and tested for different demand conditions with varying mean and variance and the response variables like WIP at different workstations and throughput were monitored. The ranges for both mean and standard deviation of the demand patterns tested lay between 0 and 10 items/hr. Based on the analyses, four different distinct state behaviors were observed (Figure 3) – static (light purple), steady & low amplitude (yellow), steady & high amplitude (green), and unsteady (dark purple). It is also observed that behaviors found to be more sensitive to variations than the level of demand (mean).

The results presented in Figure 3 motivates the dangers in defining single operation policies, the static (Type 1) behavior of throughput can be observed across a band where both the mean and variance are increased proportionally. However, the system gradually starts to shift to steady (Types 2 and 3) before finally going to unsteady state (Type 4) with increasing variance of demand for the same mean. Also, increasing the mean keeping the variance constant will shift the pattern from static to steady. Similar conclusions can be drawn

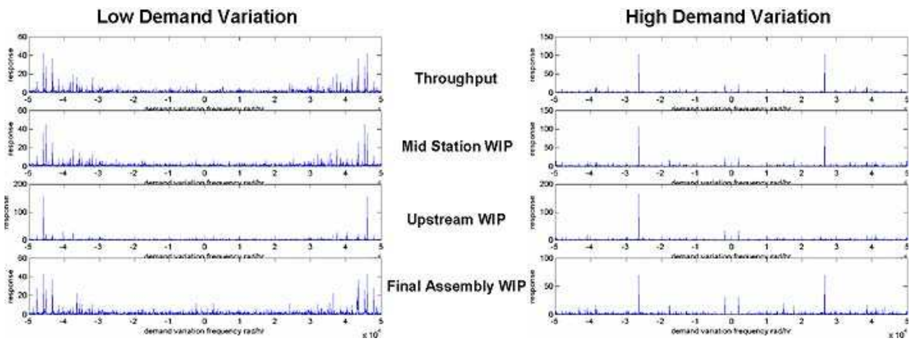


Figure 4. Frequency Response Plots for the manufacturing system.

from the other WIP plots; the variance in the demand is essentially causing the instability in the system. From the above discussion, one can conclude that the inventory and other replenishment policies have to be carefully designed by considering the relative location of the system response on the map.

b. Frequency Domain Analysis:

Frequency response analysis is important for three primary reasons. First, if the frequency response is known the response of the manufacturing system (throughput or WIP) can be predicted for any input conditions. Sinusoidal waveforms can be combined to form other (non-sinusoidal) waveforms and can be used to understand the system's response to more complex inputs. Second, by understanding this behavior one can design different manufacturing strategies for different operating conditions, with particular frequency characteristics. Third main advantage in analyzing in frequency domain is that nonlinear effects, such as powers and product interactions of the parameters, can be detected with no additional experimentation. When the parameters are driven by sinusoidal oscillations, it turns out that the interactions of the parameters are driven by compound oscillations which are the sums of sinusoidal oscillations. Figure 4 shows the frequency response plots for both low demand and high demand variation case, as you increase the variation associated with the incoming demand, the dominant frequencies become distinct. For the case of high demand variation, the dominant frequency is about 2.75×104 rad/hr (458 rad/sec) i.e., in other words if the frequency of the incoming demand is 2.75×104 rad/hr the manufacturing system exhibits unstable behavior.

c. Linearized Response Analysis:

As evidenced from Figure 3, the system response under varying conditions of demand is nonlinear. In order to fully understand this system behavior, the nonlinear response is linearized around a particular operating settings. The

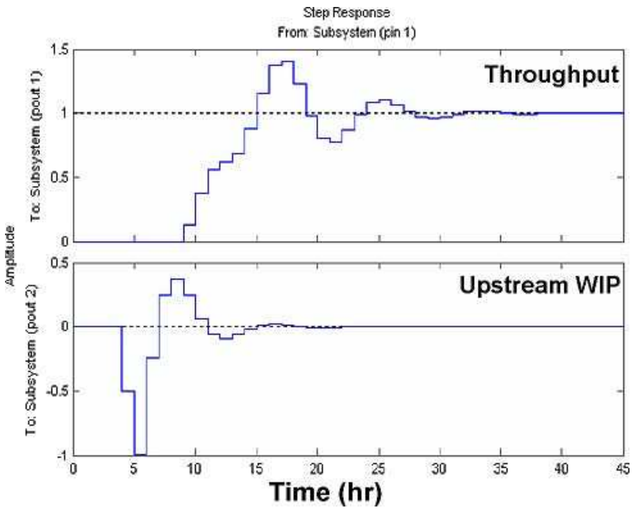


Figure 5. Step Response for Products delivered (top) and WIP3 in response to a unit demand.

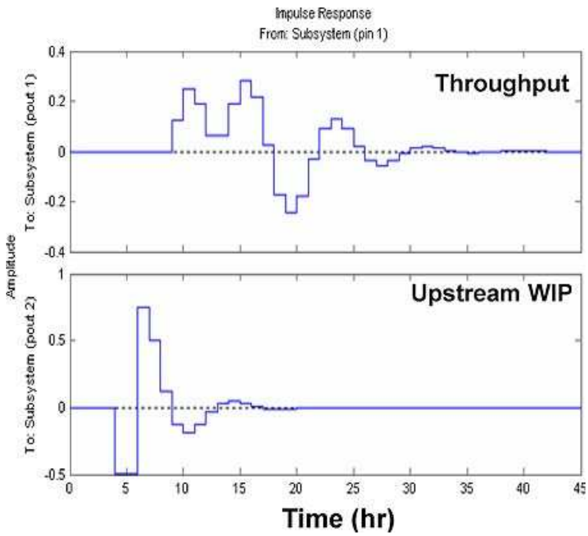


Figure 6. Impulse Response of Throughputs and WIP (Station 3).

step response for this linearized model is shown in Figure 5 and Figure 6. Figure 5 shows the linearized response for the throughput and work-in-process at workstation 3 (WIP3) for a system with no initial buffer between the workstations. As indicated the total lead time for this system is 12 hours, and it takes about 25 hours (13 hours after the impact is felt) for the system to settle down and to recover from the impact with no steady state error. During the transient phase of products delivered, it appears that there are some backlogs in the orders. The WIP3 step response is more damped than the products delivered response. The inventory or buffer policies between the workstations determine the damped nature of the response, for example Figure 5 represents the linearized step response with no initial buffer is a critically damped system, while the response in Figure 6 is an under-damped system. Based on the desired objectives of the manufacturing system, one can design inventory control policies that are fast reacting, and more proactive in negating the effects of this impact.

4.4 Comparison with Conventional Approaches

The model developed in the previous section assumes a thorough understanding of the physical phenomenon and relationships of a process. However, most often this is not a viable option. Traditional approaches using Laplace transforms (and other tools enunciated as part of control theory) or min-max-plus algebraic approaches (discrete event models) are based on the formulations obtained based on linear assumptions. Our proposed approach (next section) addresses these issues by formulating a nonlinear dynamic model from real time data gathered from the shop floor.

The use of DES models (since the states are simply discrete with no metric structure) can cause combinatorial explosion of the state space. For e.g., if there are x1 and x2 variables that describe the system, each can take the value of 0 or 1. Then the state space is only 4. On other hand, if there are 20 variables, x1 ... x20, each can only take 0 or 1. The state space is 220. This is the combinatorial or exponential explosion will lead to large computational complexity, as state transitions are usually not compactly representable using analytical functions

It was observed from our experience and from the literature that the traditional discrete event based simulations take longer time to converge to a solution, compared to continuous time based simulations. The discrete event based models take about 20 hours to complete 100 simulations of 20 years while system dynamics model can accomplish a similar task in about 1 hour (Johnson, 2002). This result is very important because it does have implications on how and what kinds of models need to be used to investigate specific problems.

Scholz-Reiter et al. (Scholz-Reiter, 2005) showed that, for a chosen arrival pattern, the autonomous control strategies can reduce throughput times and are robust against unexpected disturbances. The use of DES models allows a good description of real-world shop floor processes, but implementing autonomous control strategies involves high programming effort. In contrast, a continuous dynamic model allows an easy implementation of autonomous controls, although it describes logistic processes at a higher aggregation level compared to a DES model.

4.5 Future Modeling Approaches

As mentioned earlier, modern manufacturing plants operate in a very data rich environment. The data manifests from different sources and devices that are designed for different specific purposes like quality control, safety control, process control etc. The data from these different sources on a shop floor may be harvested into information by building real time dynamic models of the underlying system. The proposed modeling approach tries to apply dynamic systems principles to derive data-driven models for manufacturing systems. Nonlinear dynamic systems theory provides tools for modeling this complexity and allows continuous monitoring and prediction of dynamic behavior during operations.

5. Summary

This paper presented and critically compared alternative modeling and monitoring approaches for studying the performance of a production system at a process or a machine level as well as at the system level. From the foregoing, it is evident that principles and practice of manufacturing can be considerably advanced from using nonlinear dynamics and chaos theory. The key reason is that the dynamics of most manufacturing processes and systems is nonlinear. While the research at the process and machine levels have led to the development of advanced monitoring and control approaches, challenges in characterizing and modeling systems complex behaviors pose are affecting the development of adequate predictive modeling approaches for manufacturing enterprise systems.

Based on a continuous model developed for a simple manufacturing system using production control principles, control theoretic analysis were conducted to gain insight into the performance of the system, in order to better predict the system performance under different conditions of both external and internal disturbances. These dynamic models and results obtained can be used to better design/redesign the production control policies to accommodate certain deficiencies in the performance.

The research efforts thus far have addressed the adaptation of the existing results in nonlinear dynamics body of knowledge into manufacturing. We feel that the future efforts must address advancement of nonlinear dynamics principles, as well as integration of the existing methods with other analysis tools in order to realize the full potential of nonlinear dynamics and chaos theory for manufacturing. For example, nonlinear dynamics researchers have traditionally not addressed the issues of how to construct nonlinear models from actual process data, and how to treat noise present in real-world processes. However, these issues are of great importance to manufacturing process and systems characterization, especially since adequate and reliable data is very difficult to obtain in manufacturing systems (Whitehouse, 1994). Research efforts addressing these problems will advance not only manufacturing principles and practice, but also will render a fundamental contribution to the field of nonlinear dynamics.

Acknowledgments

The authors thank the National Science Foundation (DMI 0428356) for support the research presented in this book chapter. The authors acknowledge the support of their many current and past students without whose efforts this work would not have been possible. In specific, the efforts of Mr. Shankar Earni in conducting the study on manufacturing system dynamics are acknowledged. Also the authors wish to thank Mr. Jayjeet Govardhan for his help with formatting the document.

References

H. Nayfeh and B. Balachandran, *Applied Nonlinear Dynamics: Analytical, Computational and Experimental Methods*. New York, NY: Wiley, 1995.

F. Moon, *Chaotic Vibrations*. New York, NY: Wiley, 1987.

S. A. Tobias, *Machine Tool Vibration*. New York, NY: Wiley, 1965.

F. Koenigsberger and J. Tlusty, *Machine Tool Structures*. Oxford, UK: Oxford University Press, 1970.

K. F. Ehmann, S. G. Kapoor, R. E. DeVor, and I. Lazoglu, "Machining process modeling: A review," *ASME Journal of Manufacturing Science and Engineering*, vol. 119, pp. 655–663, 1997.

P. Albercht, "Dynamics of the metal cutting process," *ASME Journal of Engineering for Industry*, vol. 87, pp. 429–441, 1965.

E. Minis, E. B. Magrab, and I. O. Pandelidis, "Improved methods for the prediction of chatter in turning, Part III: A generalized linear theory," *ASME Journal of Engineering for Industry*, vol. 112, pp. 28–35, 1990.

J. Tlusty and F. Ismail, "Basic non-linearity in machining chatter," *CIRP Annals*, vol. 30, pp. 299–304, 1981.

Grabec, "Chaotic dynamics of the cutting process," *International Journal of Machine Tools and Manufacturing*, vol. 28, pp. 19–32, 1988.

J. S. Lin and C. I. Weng, "Nonlinear dynamics of cutting process," *International Journal of Mechanical Sciences*, vol. 33, pp. 645–657, 1991.

R. Komanduri and R. H. Brown, "On the mechanics of chip segmentation," *ASME Journal of Engineering for Industry*, vol. 103, pp. 33–51, 1981.

S. T. S. Bukkapatnam, A. Lakhtakia, and S. R. T. Kumara, "Analysis of sensor signals shows that turning process on a lathe exhibits low-dimensional chaos," *Physical Review E*, vol. 52, pp. 2375–2387, 1995.

H. Nayfeh, C. M. Chin, and J. Pratt, "Perturbation Methods in Nonlinear Dynamics – Applications to Machining Dynamics," *ASME Journal of Manufacturing Science and Engineering*, vol. 119, pp. 485–493, 1997.

T. Kalmar-Nagy, J. R. Pratt, M. A. Davies, and M. D. Kennedy, "Experimental and analytical investigation of subcritical instability in metal cutting," Proceedings of 17th ASME Biennial Conference on Mechanical Vibration and Noise, Las Vegas, NV, 1999.

S. T. S. Bukkapatnam, S. R. T. Kumara, and A. Lakhtakia, "Sensor-based real-time nonlinear chatter control in machining," *CIRP Journal of Manufacturing Systems*, 1999.

S. Bukkapatnam and R. Palanna, "Experimental characterization of nonlinear dynamics of grinding process," *ASME Transactions Journal of Manufacturing Science and Engineering*, 2005.

S. T. S. Bukkapatnam, S. R. T. Kumara, and A. Lakhtakia, "Dependence of computed trajectories on the step size in a nonlinear dynamic system: An investigation into cutting tool dynamics," *IIE Transactions*, vol. 27, pp. 519–529, 1995.

S. T. S. Bukkapatnam, S. R. T. Kumara, and A. Lakhtakia, "Analysis of acoustic emission in machining," *ASME Journal of Manufacturing Science and Engineering*, vol. 121, pp. 568–573, 1999.

S. T. S. Bukkapatnam, S. R. T. Kumara, and A. Lakhtakia, "Fractal estimation of flank wear in machining," *ASME Journal of Dynamic Systems, Measurements and Control*, vol. 122, pp. 89–94, 2000.

S. R. T. Kumara, S. V. Kamarthi, S. T. S. Bukkapatnam, and J. W. Lee, "Sensor based monitoring for real-time quality control in manufacturing," Proceedings of AAAI-94 Spring Symposium, 1994.

J. C. Schouten, F. Takens, and C. M. van den Bleck, "Estiamtion of the fractal dimension of a noisy attractor," *Physical Review E*, vol. 50, pp. 1851–1861, 1994.

I. M. Johnstone and B. W. Silverman, "Wavelet threshold estimators for data with correlated noise," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 59, pp. 1111–1467, 1997.

S. V. Kamarthi, "On-line flank wear estimation in turning using multi-sensor fusion and neural networks," in *Department of Industrial and Manufacturing Engineering,* PhD Thesis. University Park, PA: The Pennsylvania State University, 1994.

G. Byrne, D. Dornfeld, I. Inasaki, G. Ketteler, W. Konig, and R. Teti, "Tool condition monitoring (TCM) – The status of research and industrial application," *CIRP Annals*, vol. 44, pp. 541–567, 1995.

T. Blum and I. Inasaki, "A study on acoustic emission from the orthogonal cutting process," *ASME Journal of Engineering for Industry*, vol. 112, pp. 203–211, 1990.

E. Kannatey-Asibu and D. A. Dornfeld, "Quantitative relationships for acoustic emission from orthogonal metal cutting," *ASME Journal of Engineering for Industry*, vol. 103, pp. 330–340, 1981.

J. J. Liu and D. A. Dornfeld, "Modeling and analysis of acoustic emission in diamond turning," *ASME Journal of Manufacturing Science and Engineering*, vol. 118, pp. 199–207, 1996.

H. Knapp, "Suboptimal wavelet decompositions for near-periodic signals," in *Mathematics Department*, M.S. Thesis. LA: University of Southern California, 1998.

S. R. T. Kumara, J. Suh, and S. P. Mysore, "Machining fault diagnosis and prognosis: Application of advanced signal processing technique," *CIRP Annals*, vol. 48, pp. 317–320, 1999.

B. Bhushan and G. Ganti, "Generalized fractal analysis and its applications to engineering surfaces," *Wear*, vol. 180, pp. 17–34, 1995.

D. J. Whitehouse, D. K. Bowen, V. C. Venkatesh, P. Leonardo, and C. A. Brown, "Gloss and Surface Topography," *CIRP Annals*, vol. 2, pp. 1–9, 1994.

C. A. Brown, W. A. Johnsen, and K. M. Hult, “Scale-Sensitivity, Fractal Analysis and Simulations,” *International Journal of Machine Tools and Manufacture*, vol. 38, pp. 633–637, 1998.

C. A. Brown, P. D. Charles, W. A. Johnsen, and S. Chesters, “Fractal Analysis of Topographic Data by the Patchwork Method,” *Wear*, vol. 161, pp. 61–67, 1993.

G. A. Stark and K. S. Moon, “Modeling surface texture in the peripheral milling Process Using Neural Network, Spline, and Fractal Methods with Evidence of Chaos,” *Journal of Manufacturing Science and Engineering*, vol. 121, pp. 251, 1999.

S. Earni and S. T. S. Bukkapatnam, “Towards performance analysis of manufacturing systems using nonlinear dynamics principles,” Proceedings of SAE, March 2006.

J. D. Sterman, “Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment,” *Management Science*, vol. 35, pp. 321–339, 1989.

A. V. Ackere, E. R. Larsen, and J. D. W. Morecroft, “Systems thinking and business process redesign: An application to the Beer Game,” *European Management Journal*, vol. 11, pp. 412–423, 1993.

P. Kaminsky and D. Simchi-Levi, “A new computerized Beer Game: A tool for teaching the value of integrated supply chain management,” *Production and Operations Management Society in POMS Series in Technology and Operations Management “Global supply chain and technology management”*, vol. 1, pp. 216–225, 1998.

J. Forrester, *Industrial dynamics*. Cambridge MA: MIT Press, 1961.

D. R. Towill and D. Vecchio, “The application of filter theory to the study of supply chain dynamics,” *Production Planning & Control*, vol. 5, pp. 82–96, 1994.

D. R. Towill, “Industrial dynamics modelling of supply chains,” *International Journal of Physical Distribution & Logistics*, vol. 26, pp. 23–42, 1993.

J. Dejonckheere, S. Disney, M. Lambrecht, and D. R. Towill, “Transfer function analysis of forecasting inducted Bullwhip in supply chains,” Eleventh International Working Conference on Production Economics, Igls, Austria, 2000.

F. Chen, Z. Drezner, J. K. Ryan, and D. Simchi-Levi, “Quantifying the Bullwhip effect in a simple supply chain: The impact of forecasting, lead-times and information,” *Management Science*, vol. 46, 2000.

F. Chen, J. K. Ryan, and D. Simchi-Levi, “The impact of exponential smoothing forecasts on the bullwhip effect,” *Naval Research Logistics*, vol. 47, pp. 269–286, 2000.

K. Popplewell and M. C. Bonney, “The application of discrete linear control theory to the analysis of multiproduct, multi-level production control systems,” *International Journal of Production Research*, vol. 25, pp. 45–56, 1987.

R. G. Coyle, “Assessing the controllability of a production and raw materials system,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-12, pp. 867–876, November/December 1982.

G. Cachon and M. Fisher, “Campbell soup’s continuous replenishment program: evaluation and enhanced inventory decision rules,” *Productions and Operations Management*, vol. 6, pp. 266-276, 1997.

G. Pritschow and H. P. Wiendahl, “Application of control theory for production logistics – results of a joint project,” *CIRP Annals*, vol. 44, pp. 421–424, 1995.

J. Kilpatrick, *Lean Principles*: Utah Manufacturing Extension Partnership, 2003.

V. Parunak, “Characterising the manufacturing scheduling problem,” *Journal of Manufacturing Systems*, vol. 10, pp. 241–258, 1991.

H. P. Wiendahl and H. Scheffczyk, “Simulation based analysis of complex production systems with methods of nonlinear dynamics,” *CIRP Annals*, pp. 357–360, 1999.

L. Bongaerts, P. Valckenaers, H. V. Brussel, and P. Peeters, "Schedule execution in holonic manufacturing systems," Proceedings of XXIX CIRP International Seminar on Manufacturing Systems, Osaka, Japan, 1997.

B. Scholz-Reiter and S. Mueller, "Throughput time control in production systems supported by neural networks," *CIRP Annals*, vol. 49, pp. 331–334, 2000.

H.-P. Wiendahl and J. Breithaupt, "Modelling and Controlling the Dynamics of Production Systems," *International Journal of Production Planning & Control*, vol. 10, pp. 389–401, 1999.

J. H. Kim and N. A. Duffie, "Backlog control design for a closed loop PPC system," *CIRP Annals* vol. 53, pp. 357–360, 2004.

A. M. Ratering and N. A. Duffie, "Design and analysis of a closed-loop single-workstation PPC system," *CIRP Annals*, vol. 52, pp. 355–358, 2003.

H. H. Wiendahl, N. Roth, and E. Westkamper, "Logistical positioning in a turbulent environment," *CIRP Annals*, vol. 51, pp. 383–386, 2002.

S. Cho and V. V. Prabhu, "Sliding mode dynamics in continuous feedback control for distributed discrete-event scheduling," *Automatica*, vol. 38, pp. 1499–1515, September, 2002.

N. Duffie and I. Falu, "Control-theoretic analysis of a closed-loop PPC system," *CIRP Annals*, vol. 51, pp. 379–382, 2002.

S. Johnson and B. Eberlein, "Alternative Modeling Approaches: A Case Study in the Oil & Gas Industry," Conference of the System Dynamics Society, 2002.

B. Scholz-Reiter, M. Freitag, C. d. Beer, and T. Jagalski, "Modeling Dynamics of Autonomous Logistic Processes: Discrete-event versus Continuous Approaches," *CIRP Annals*, vol. STC O-54, pp. 413, 2005.

Chapter 6

Evolutionary Traffic Flow Landscapes: A Fitness Approach for ITS Management

Kingsley E. Haynes,[1] Rajendra G. Kulkarni,[2] and Roger R. Stough[3]

[1] *The School of Public Policy, George Mason University, Fairfax, VA 22030, E-mail: khaynes@gmu.edu*

[2] *Center for Regional Analysis, The School of Public Policy, George Mason University, Fairfax, VA 22030*

[3] *The School of Public Policy, George Mason University, Fairfax, VA 22030*

Abstract The road patterns of major metropolitan areas and constituent jurisdictions evolve slowly through a complex set of independent and interdependent decisions producing a transportation network. The resulting network must be used for variety of commuting and spatial interaction activity. A typical trip taker spends considerable time on the road to reach the workplace and other destinations. Adding more links to existing road networks and/or increasing traffic capacity by adding lanes does not necessarily decrease travel times (e.g. Braess' paradox). However a dense redundant network of roads provides a trip taker with alternate routes when traffic jams occur. Such issues raise the question of, how to evaluate the flow characteristics of the entire road network of a jurisdiction and its larger region? How might the impact of adding more links/lanes or blocking existing links/lanes be best measured?

To answer these and related questions, we propose a methodology to evaluate a fitness criteria for road networks based on Kauffman's NK model (1993) and develop an information theoretic measure of the order or organization in transportation networks.

Keywords: fitness landscapes; NK model; entropy; organization; ITS technology

1. Introduction

Urban road networks are characterized by traffic congestion, incidents and accidents (Lave, 1985), resulting in travel delays for commuters and other trip takers on urban road networks (Downs, 1992). The interaction costs of such congestion in a regional economy is enormous and factoring in work time lost to business and commuters makes the sums astronomical (Arnott, Small, 1994). Increasing capacity of existing freeways by adding more lanes is not always possible or environmentally desirable and does not always ease the delays. The much studied Braess's paradox tells us that congestion may increase as capacity is increased instead of reducing it (Murchland, 1970). However the costs of incidents and accidents could be reduced if the trip taker is provided with timely warnings of such events. ITS (Intelligent Transportation Systems) traveler management systems hold the promise of providing information on traffic conditions. However, providing data on traffic conditions alone may be of little help if there is no underlying processing framework to evaluate and disseminate the processed information. So what kind of framework might be useful to process, evaluate and disseminate network flow data? Figures 1 and 2 show schematics of a possible framework. An urban traffic flow network divided among a number of zones has one or more traffic management center (TMC) in each zone. The data collected by non-intrusive surveillance equipment (Figure 2) in each zone is processed by a traffic management center (TMC) in that zone and disseminated to users. In this paper we develop an analytical model of the TMC data processing unit and suggest some underlying considerations that need to be assessed in developing a decision framework for traffic guidance and management. The analytical model is based on concepts borrowed from evolutionary biology, especially the concept of fitness landscapes (Kauffman, 1987) and information theory (Appelbaum, 1996; Suhir, 1997) to describe the organization or order of traffic networks.

So far, with the exception of a few (Herman, 1982) almost all of the modeling efforts to describe network traffic flows are based on physical analogs. A survey of the literature shows many attempts to model the dynamics as well as the equilibrium/disequilibrium network flow conditions that exist on urban road networks. Both analytical and simulation/experimental studies have been carried out (Friesz, Bernstein and Stough, 1996; Friesz, Bernstein, Smith, Tobin and Wie, 1993; Friesz, Bernstein, Mehta, Tobin and Ganjalizadeh, 1994; Mahmassani, 1995; Mahmassani, 1990; Mahmassani, Hu and Jayakrishnan, 1992; Mahmassani and Peeta, 1992; Koutsopoulos, 1995.) While, some modelers have addressed the stochasticity of traffic flows by trying to reduce the randomness – following the so called micro-simulations approach (for pioneering work see Chang, Mahmassani, Herman, 1985; Mahmassani, Chang, 1987) and the TRANSIMS model developed by the transportation group at Los Alamos National Laboratory (Berkbigler, Loose, Davis, Williams, 1995; Smith, Beckman, Anson, Nagel, Williams, 1995).

In this paper we look at urban traffic flow networks as *open dynamic systems* consisting of a large number of agents (users) who interact with each other and with changes in the environment, similar to biological agents who interact and

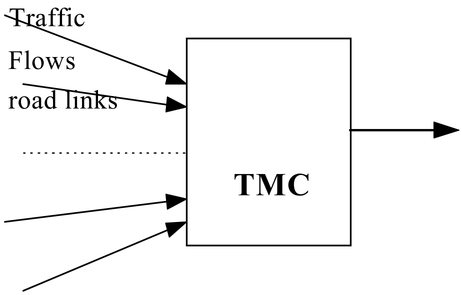


Figure 1. Input data on traffic flow.

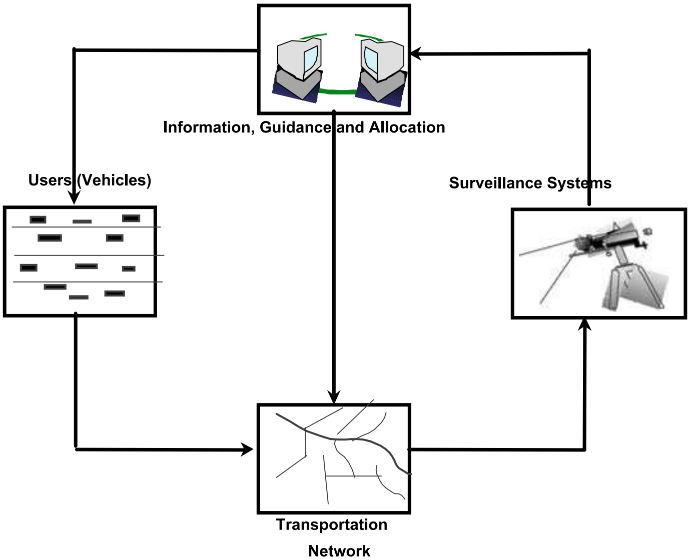


Figure 2. A Traffic Management Center (TMC) Flow Chart (schematic inspired by Mahmassani, 1995).

adapt to ever changing surroundings. In such cases it may be that biological and information theory based analogs rather than traditional physical analogs could be used to explain complex dynamic traffic flow systems. Of course, such an analogy to biological systems is limited to those evolutionary properties that help explain traffic flows. Next we explain the motivational aspects for considering a biological analog of a network traffic flow model.

2. Motivation

In various degrees and shades biological models have been adapted by fields as different as cosmology (evolutionary universe; Linde, 1994; Coleman, Hartle, Piran, Weinberg, 1991), economics (evolutionary economics; Tu,1992; Krugman, 1994, 1995; Arthur, 1989), and sociobiology (Wilson, 1995), to name a few. Although many workers in other fields view evolution as a concept to describe gradual changes (as opposed to revolutionary changes) in system behavior (Fabian, 1998), this perspective is not necessarily consistent with the evidence of punctuated evolution. Whatever the points of view, all evolutionary systems consist of a large number of agents whose interactions give rise to complex system-wide behavior: micro-level actions giving rise to macro-level patterns of behavior.

Consider for example, an urban road traffic network consisting of a large number of road segments. Workday traffic on a segment of a highway, typically has the profile of the morning and evening rush hour peaks with the intervening troughs for the rest of the day. However, it is quite unlikely that the traffic pattern profile on a given day matches exactly with those of previous or following workdays. Indeed the stochasticity of traffic patterns arises in part as a consequence of "non-collaborative" trips taken by commuters. Commuters are "aware" of other commuters' plans to travel only to the extent that they are going to share the limited resources of time and road space with other unknown commuters. The commuters do not collaborate or inform each other of their intended trips and schedules and plan accordingly for their journeys. Most of the time, commuters follow a loose schedule that they create out of their day to day experiences of trips on the roads. Thus, traffic is an aggregate of the multitudes of decisions executed by commuters in a non-collaborative manner giving rise to traffic patterns. And this occurs despite of master plans intended to regulate traffic on the roads. Can information theory models help explain the traffic patterns resulting from actions of distributed agents? Do biological and information theory models help in understanding concepts such as organizational order and adaptation? How good an explanation of traffic flow networks is such a metaphor for the life-like processes.

The appeal of biological models comes from the following. Although they are somewhat *imprecise and fuzzy,* they provide a powerful explanatory framework for dealing with systems consisting of independent but interacting agents; who change their behavior to suit the dynamics of the environment. So a model based on these concepts could provide us with some insight into network traffic flows. Such an insight might be used in the development of tools to assist ITS systems and thereby improve the traffic flows in the network. For example, suppose that we know about a specific property that is beneficial to a population of agents. Then at least in theory, an abstract landscape defined in terms of this specific property can be constructed and the dynamics of this property observed as the population adapts to and modifies the changing landscape. In evolutionary biology such landscapes are defined by survival fitness. In the analytical model described in this paper, we use the *ease of flow of traffic on the network* as a fitness property to construct abstract traffic flow landscapes, similar to the N-K model fitness landscapes proposed by Kauffman (1993).

However there are some important differences between the analytical model presented here and the N-K model. The N-K model describes interactions among agents in terms of boolean functions and relies on autonomous boolean networks for emergence of organization. In this paper we use an information theoretic approach to compute the ability of networks to display organization and how such a method can be utilized by ITS to adapt to changes in network flows. Additionally, we explain how the use of ITS related technologies can help in maintaining an overall good fitness of traffic networks for the benefit of both the users and the traffic managers.

The following part (III) describes the background of the current effort. It includes a brief description of how traffic flows are measured, followed by an introduction to the concept of fitness landscapes a` la the NK model. A brief explanation of a network traffic flow landscapes is outlined. Section 4 develops the analytical traffic flow landscape model, with a few rudimentary examples. Part V describes a search process to find the higher fitness configurations of a traffic flow landscape. Part VI specifies an information theoretic approach to organization of traffic flow landscapes. Conclusions and future work are discussed in the final part (VII).

3. Background

An urban region's road network consists of many types of roads – highways, major and minor roads, arterials and connecting roads. For a traffic fitness landscape only those roads and links that are referred to as primary and secondary roads/links as described in the TIGER/Line™ files Census Feature Class Codes (1992) are included. Links are segments on highways, major roads and arterials. Segments are characterized by the levels of service also called an LOS (Highway Capacity Manual, 1985) which varies depending on the traffic conditions on these segments. Thus for example, levels of service 'A' through 'F' are the six possible LOS. According to the Special Report 209 of the Highway Capacity Manual, "the concept of levels of service is defined as a qualitative measure describing operational conditions within a traffic stream, and their perception by motorists and/or passengers. A level-of-service (LOS) definition generally describes these conditions in terms of such factors as speed and travel time, freedom of maneuver, traffic interruptions, comfort and convenience, and safety" (1985, pp. 1–3).

Fitness Criteria For Traffic Flows

Suppose we assign a numerical value to each LOS, then in theory we could use these values as a fitness measure of each segment. Unlike the concept of fitness in biology which is a property associated with survivability of a species, the fitness in the traffic flow model refers to the idea of relative ease (high fitness) or difficulty (lower fitness) of flow of traffic on segments of roads in transportation network. The point is that, one can use the LOS concept to represent traffic flow conditions at any time over an entire road transport network. To quantify the qualitative concept of levels of service we suggest a very simple method in Appendix A based on fractals.

Network Traffic Flow Landscapes

The word "landscapes" has topological connotations. Even though, the word evokes images that everyone relates to in different ways, there are certain properties that are common to many landscapes. For instance the landscapes have morphology such as multiple peaks (either sharp or gentle) and troughs and connecting ridges. The topology of a region makes it clear that to reach point 'P' on one of the peaks from point 'Q' on another peak involves finding the best possible route between these two points, avoiding regions of valleys. Alternately, one may want to avoid the peaks and reach a desired valley region. The landscape image suggests that there may be more than one peak that may satisfy, given a set of criteria and that not all peaks and valleys are reachable easily. Thus one may envision assigning a fitness value to criterion or criteria set and then creating a visual image with peaks for good fitness values and valleys for bad fitness. So how does one use the idea of fitness to create traffic network landscape?

A network of 'N' segments with 'L' number of LOS has L^N possible configurations. (In terms of levels of service A through F, we have 6^N possible configurations for N segment network.) Each configuration is just one LOS different than its neighboring $D = N*(L - 1)$ other configurations. As was mentioned earlier we associate with each LOS of a segment, a computed fitness value. Note that, LOS of a segment may affect the LOS of its neighboring segments. In the following part, we suggest an analytical expression to take these interactions into account. The number of such interacting segments is represented by an integer 'K', called the interaction parameter. Let us assume for simplicity that, the fitness values of each segment are additive. Then, a configuration where all segments of the network have LOS 'A', has the highest fitness value (free flow traffic), say 'Alpha'. At the other end is a configuration

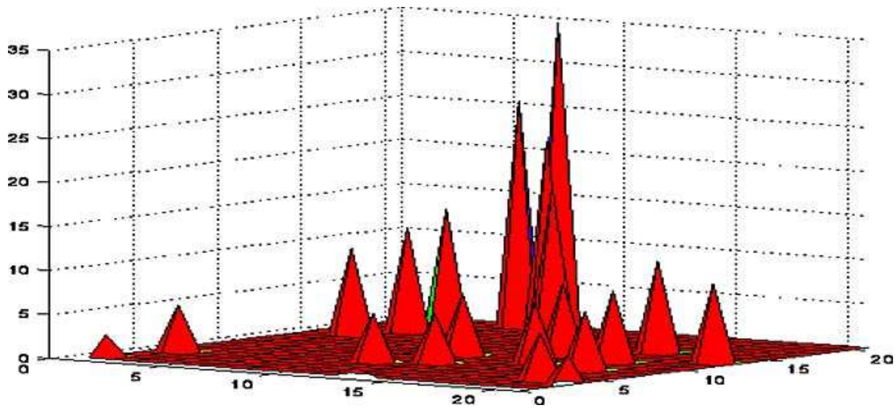


Figure 3. Schematic of a hypothetical fitness landscape.

where every segment has a LOS of ‘F’, this certainly has the lowest fitness value (traffic jams on all segments), say ‘Omega’. Usually, the network with ‘N’ segments will be in one of the L^N configurations with a value of fitness that is less than ‘Alpha’ and more than ‘Omega’. But, the overall fitness of the network is not a simple additive process, since the traffic on each segment is affected by other segments, and more by near segments (nearest neighboring) than far segments. Hence each segment’s fitness is a function of the fitness of neighboring segments. Let us assume that there are ‘K’ such neighbors that affect the fitness of any segment, where the value of ‘K’ can be between zero and ‘N – 1’. When K = 0, each segment has traffic flows that are independent of all other segments. On the other hand, when K = N – 1, each segment’s traffic flow is affected by the traffic flows on all other segments.

Description of traffic networks in terms of L^N configuration is analogous to a system with L^N states. A truly random system would show ergodic behavior such that the probability of such a system being in any one of these states is the same. But are traffic flows truly random? Probably not. Since traffic systems show patterns at the macro level, we need to assess the degree to which these patterns are constant or degree to which they evolve and organize into new patterns that are mostly beneficial to participating agents. In the following section we illustrate the development of the analytical model by a simple example, almost a cartoon of a real life network.

4. Analytical Model of Traffic Flow Landscape

Consider an urban road network of N segments. The segments can be either sections between mile stones or distances between consecutive traffic signals or any other measure that has been defined with consistency across the network.

The traffic flow levels on each segment are determined in terms of the LOS. Thus each segment can have ‘A’ through ‘F’ levels of service. As a simplified illustration, let us restrict the flow conditions to two levels of service, LOS of ‘A’ for free flow and LOS of ‘F’ for no flow. Let us assume that the fitness contribution of a segment with LOS of ‘A’ is +1.0 and that with LOS of ‘F’ is 0.0. Obviously, it is possible to use another set of criteria to assign a different set of fitness values. At any instant a set of segments with specific flow levels constitutes a configuration of the entire road network among all possible configurations. For example, for ‘N’ segments with ‘L’ flow levels, the number of possible configurations is L^N. For two types of flows (identified by A and F) the total number of possible configurations of segment flows is 2^N. Let ‘K’ refer to number of segments interacting with each other. Next we explain the road configurations for different values of ‘K’.

4.1 K = 0 Case

Let us assume that a segment with full flow condition (A) contributes +1.0 to the overall fitness of the network, and a blocked segment (F) contributes 0.0 to the overall fitness of the network. Note that in this case, ‘K’ the number of interactions among the segments is assumed to be zero, i.e., each segment contributes to the overall fitness independent of all other segments. The entire network configuration may be represented as a combination of +1.0 and 0.0s. The total number of possible states is 2^N. Thus, there are two trivial states that one can find for the N segment network. One of these states has all the segments blocked thus the network segments are represented as the following vector:

$$(F_1, F_2, F_3, \ldots, F_N), \tag{1}$$

and the total fitness M_g, is given by the following:

$$M_g = \sum_{i=1}^{N} f(F_i) = 0.0. \tag{2}$$

The other state has all the segments in the free flow condition and may be represented as the following vector:

$$(A_1, A_2, A_3, \ldots, A_N), \tag{3}$$

and the total fitness M_g, for this state is given by the following:

$$M_g = \sum_{i=1}^{N} f(A_i) = +N. \tag{4}$$

While every other state has a total fitness contribution that is between 0.0 and N. Thus all the network states with their fitness contributions can be represented in the form of an N*(L – 1) = D dimensional hypercube where each vertex represents a state or configuration of the network and the value assigned to the vertex is the fitness of the network in that particular configuration. Every configuration is one LOS different than its neighbors in a 'D' dimensional hypercube. This hypercube is the simplest example of a fitness landscape, it has one minimum and one maximum and the rest of the fitness values are between 0 and N. For example, if N = 4 and L = 2, then, the total number of combinations are $L^N = 2^4 = 16$. Table 1 shows each configuration and its total fitness contribution. It is clear from the fitness values that there are multiple configurations with the same fitness value. Configurations 2, 3, 5 and 9 all have fitness of 1/4, while configurations 8, 12, 14 and 15 have fitness of 3/4. There are 6 configurations with 1/2 fitness (4, 6, 7, 10, 11, 13). Thus if we assign probabilities to each of these fitness values then it is clear that the configuration with fitness value of 1/2 has a higher probability of occurrence. And intuitively it makes sense that a configuration with all blocked segments has a very small probability as does the configuration with all segments in a free flow situation. Thus even a very simple coding scheme gives quite a bit of information on the condition of a traffic network. This schema can be extended to a greater number of levels or LOS to achieve a more realistic model of traffic flow conditions.

Table 1. Configuration vs. Fitness

Configuration	Fitness value	Avg. Fitness
		K=0
1=F,F,F,F	0,0,0,0	0
2=F,F,F,A	0,0,0,1	1/4
3=F,F,A,F	0,0,1,0	1/4
4=F,F,A,A	0,0,1,1	1/2
5=F,A,F,F	0,1,0,0	1/4
6=F,A,F,A	0,1,0,1	1/2
7=F,A,A,F	0,1,1,0	1/2
8=F,A,A,A	0,1,1,1	3/4
9=A,F,F,F	1,0,0,0	1/4
10=A,F,F,A	1,0,0,1	1/2
11=A,F,A,F	1,0,1,0	1/2
12=A,F,A,A	1,0,1,1	3/4
13=A,A,F,F	1,1,0,0	1/2
14=A,A,F,A	1,1,0,1	3/4
15=A,A,A,F	1,1,1,0	3/4
16=A,A,A,A	1,1,1,1	1

4.2 K = N – 1

Now, let us consider the case when each segment interacts with all the other segments of the network. Since, we do not yet know how each segment affects the other segments, let us assume that the complex interactions are multiplicative in nature, that is if segments have high fitness values then the result of interaction would be a high fitness contribution value. On the other hand if one or more segments has lower fitness values then the result of the interactions accordingly would reflect a low fitness contribution value.

Another way to represent the interactions is to use a modified Tanner function (Tanner, 1961; Paelinck and Klassen, 1979),

$$\pi(f_i') = \sum_{j=1}^{K+1} \pi(f_j) * \left((d_{ij} - 1)/d_{ij}\right) \times \exp(-\alpha \times d_{ij}) \qquad (5)$$

where, 'α' is a proportionality constant, 'd_{ij}' is the distance (steps) between segment 'i' and segment 'j', $\pi(f_j)$ is the fitness of segment 'j', $\pi(f_i)$ is the fitness of segment 'i' and $\pi(f_i')$ the computed fitness of segment 'i' as a result of the 'k+1' interactions between segment 'i' and other segments carried out according to equation (5).

Next let us consider a set of configurations, each with 4 segments as before. But now, let us represent the individual fitness values as random numbers between 0 and +1. Then the total fitness of a configuration Mg, is given by the following:

$$M_g = \sum_{i=1}^{N} \pi(f_i'), \qquad (6)$$

where 'g' is a configuration, $\pi(f_i')$ is the fitness potential of a segment 'i' calculated using equation (5).

Once more, the fitness landscape is constructed as an N*(L – 1) dimensional abstract hypercube where each vertex represents a configuration with a fitness value that is a contribution from the segments of that configuration. Although, the fitness landscape is bounded from above and below by 0 and N, it is now much more rugged and has multiple peaks interspersed with deep valleys. For example for a simple network of N = 4 and L = 2 (near free flow and near total blockage) we obtain a hypercube with $2^4 = 16$ configurations represented by its vertices. Each configuration on a vertex has a fitness value that is either a peak or a valley depending on the result of interactions among the four segments of each configuration. In general, every vertex has a configuration that has 'D' other neighboring vertices with their own configurations. And all of these differ from each other in one LOS of a segment. Accordingly, the total fitness also is a bit different for each vertex.

The two extreme cases of K = 0 and K = N – 1 show us how the fitness landscape can change from a one maximum and one minimum fitness landscape to a multiple peak rugged landscape (Figure 5.) Note that the landscape for the K = N – 1 case was generated using random values between 0.0 and 1.0 for fitness of each segment, and the interactions among the segments were thought to be of multiplicative type. The configurations 4, 5 and 8 have a higher fitness than the rest. The overall fitness level becomes smaller as the number of

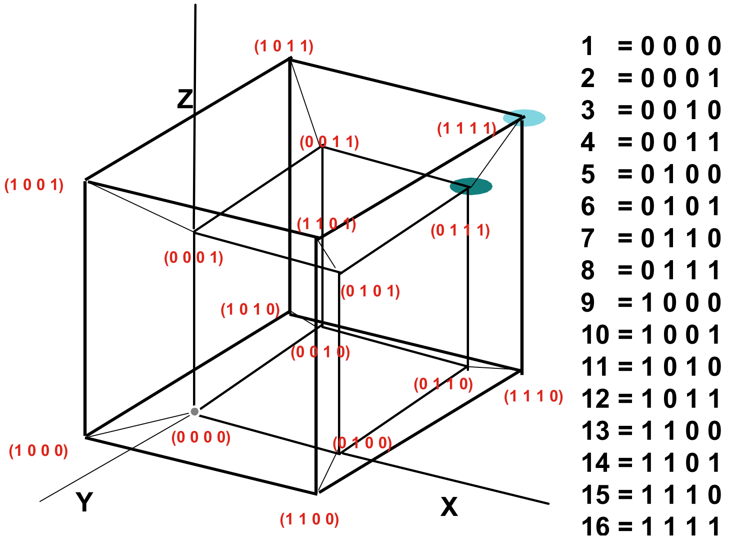


Figure 4. Hypercube Representation of Genotypes for Genes N = 4.

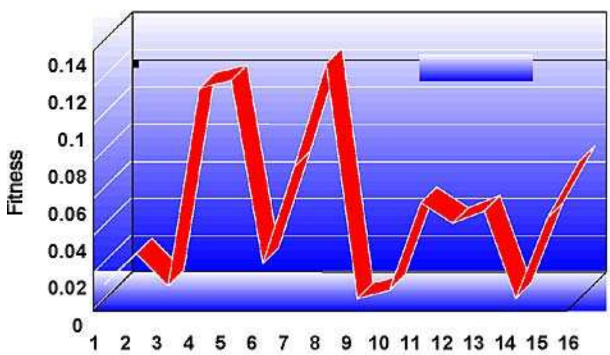


Figure 5. Fitness Landscape N = 4, K = 3, L = 2.

interactions increases from K = 0 to K = 3. This is due to the conflicting nature of interactions among the segments and results in reduced fitness maxima.

In general, it may be said that as the value of K changes the fitness landscape also changes from a single maximum/single minimum fitness landscape to a rugged multiple max/min fitness landscape whose maxima and minima have reduced fitness values. To create a fitness landscape for the entire network, one may use a random fitness function or some other function that reflects the general traffic conditions. In the later case the function could be a weighted combination of a number of traffic properties, such as peak flow times/numbers, density of the traffic, speed limits on the segments or any other relevant property of the traffic on the road network.

The specification of the traffic flow fitness landscape model is complete when the assignment of the fitness vectors for all the segments is done and the traffic flow landscape represented as a hypercube in an N*(L – 1) dimensional space for 'L' LOS traffic flow condition. As was previously mentioned, each of the configurations at the vertex is one LOS different from its 'D' neighbors and accordingly its fitness is a little bit different than the rest of its 'D' neighbors.

Now it is possible to make an estimate of the overall fitness of the road network for all configurations using equation (6) as follows:

$$\Gamma = \frac{1}{L^N} \sum_{g=1}^{L^N} M_g \tag{7}$$

where M_g is the fitness of configuration 'g'. Equation (7) serves as the general fitness index of road networks.

5. Exploring the Traffic Flow Landscape

Since the traffic flows on roads change dynamically, they do not lend themselves easily to modeling. One cannot associate a single equilibrium point at which the traffic flows settle down into a regular pattern. Instead, the traffic flows follow multi-equilibria metastable behavior, jumping from one configuration g_x to the next configuration g_y on the fitness landscape hypercube. The new configuration g_y may or may not be in the immediate neighborhood of g_x, the process of moving from this configuration to the next continues as the flow dynamics change on various segments. Note that the movement over the fitness landscapes may not always result in a better fitness configuration. Consider a configuration $g_{x,}$ from among all the other L^N configurations, then it can be shown that, in general the probability that g_x has a better fitness value than its D neighbors is given by:

$$p(g_x) = \frac{1}{D+1}. \quad (8)$$

The higher fitness value of configuration g_x makes it a locally optimal configuration among its neighbors. Then for a landscape consisting of L^N configurations, the total number of such local optima is given by:

$$E \approx \frac{L^N}{D+1}, \quad (9)$$

Thus there exist a large number of locally optimal configurations for a traffic flow landscape. The local optima are the multiple equilibria that are scattered all over the traffic fitness landscape.

Attractors and Attractor Basin

Next, let us consider the example of the simple road network with N = 4 segments, namely ‘a‘, ‘b’, ‘c’, ‘d’; and L = 2 LOS and 0 <= K <= 3 interactions. Let us rank the 16 possible configurations according to their fitness values. Let us assume that these segments are such that, during a time period ‘t’, the traffic on segment ‘a’ moves on to segment ‘b’ and so on. Let ‘0101’ be the configuration representing the current traffic conditions on the four segments. Then in one time period, as the traffic moves, the new configuration could be one of the following, ‘1010,’ the complement of previous configuration or ‘0000’ all segments congested or ‘1111,’ all segments are in free flow condition. None of these three successor configurations are one LOS differing neighbors (just one of the segments with a different flow condition) of the previous configuration. In fact the flow conditions on the successor configuration could be such that the resulting configurations are 2 or more but less than N LOS differing neighbor configurations. Next let us consider a large network with N segments. For such a network consisting of large number of segments, the successor configuration could vary from being one LOS differing neighbor to N LOS differing neighbors. If the successor configuration is the same from one instant to another or if the successor configurations change back into the original configuration then the original configuration becomes the *attractor*. In other words, if a set of different configurations corresponding to small scale perturbations in the flows on segments have the same successor configuration, then the members of the set form the so called *attractor basin* and the successor configuration may be designated as the *attractor configuration* or the so called *metastable equilibrium configuration*. On the other hand, if the successor configurations are all wildly different then it is an indication that the traffic flow patterns are changing chaotically and that the network has become unstable. The changes in the configurations from one instant to next can be measured in terms of the Hamming distance. Thus, the Hamming distance may serve as a measure of instability of a road network.

Search for Local Optima

If one could construct a traffic flow landscape at a given instant then in theory it is possible to estimate the time needed to reach an optimal solution. Suppose that a traffic landscape has been constructed and currently the entire network is represented as a configuration g_x in this landscape. Next, an incident occurs on one of the segments of the road network, the resulting traffic flow with congestion can be viewed as a configuration g_y on the same landscape and the amount of time needed for the network to reach from the current configuration corresponding to the congested segment, to one of the locally optimal equilibria points is given by:

$$T_{opt} = \sum_{t=0}^{\log_2(D-1)-1} L^t \tag{10}$$

where 'D' is the dimensionality of the traffic landscape, 'L' the number of LOS. The above result can be explained in terms of the rank ordering of the configurations, the dimensionality of the landscape and self-avoiding biased random walks on the fitness landscape. As was mentioned earlier, for a D = N*(L – 1) dimensional hypercube, a configuration at a vertex is at least one LOS different than its 'D' other neighbors. Thus if we start at a worst fitness configuration, then moving to any of its neighboring 'D' vertices would lead us to a configuration that has better fitness than the previous one. Since the total number of configurations is L^N, the rank order of the new configuration is between 2 and L^N. If one follows this in random fashion to move to a vertex that has better fitness than the previous one, then every such move makes the new configuration halfway closer to the remaining configurations. So as the improvement continues, the process slows down such that for every such move the time to search for a fitter neighbor doubles. Yet another way to look at the ranked fitness landscape starting from the worst fitness vertex g_w is similar to travelling down a tree whose root is the current vertex and at each level of the tree, each node (the new configuration) branches to 'D' other configurations (See Fig. 6.) which are 1 LOS different and accordingly have a different fitness level.

At the start of a given time period, we can construct a configuration tree that has the root configuration corresponding to the current traffic flows on the segments. Next we search randomly for a better fitness configuration on one of

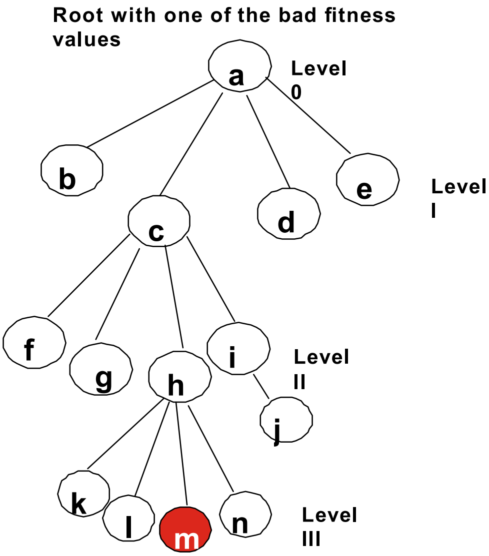


Figure 6. Partial Fitness Tree.

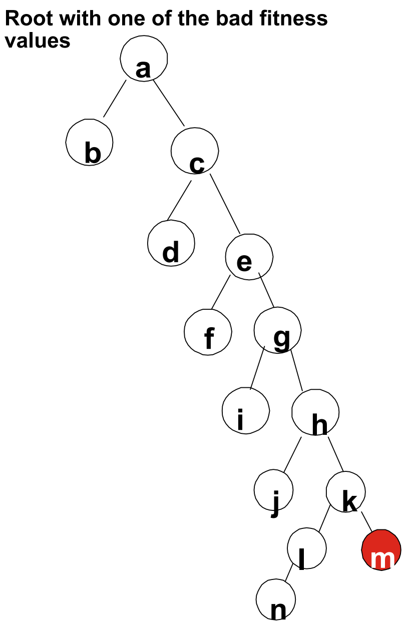


Figure 7. Partial Fitness Binary Tree.

the 'D' branches. Once such a configuration is found, then we are at the new configuration and repeat the process until we reach a local optimum. The total time or the number of levels becomes a measure of the time needed to reach a locally optimal configuration. Alternately, the configuration tree in Figure 6 can be changed into a binary tree (Horowitz and Sahni, 1985; Wilson, 1988), such that at each node or configuration there are just two branches, the right and the left branch (see Fig. 7). Every time we reach a node, we take the branch that leads to a better fitness configuration till we reach a leaf node that corresponds to a locally optimal configuration.

Similarly, we could use the above techniques to do the impact analysis on the basis of increasing (decreasing) the total number of segments (N) or changing the value of ‘K’ (the number of interactions per segment) by generation of a configuration tree and searching for a locally optimal configuration that has a better fitness than the configurations in its neighborhood.

6. Organization and Traffic Flows

Suppose that an abstract traffic fitness landscape has been constructed for ‘L’ LOS, ‘N’ segments and ‘K’ interactions. Then, from equation (6) the total fitness over all possible (L^N) configurations is given by

$$M = \sum_{g=1}^{L^N} M_g. \tag{11}$$

Even for a small number of segments the total number of possible configurations increases exponentially. For example, a network with 100 segments and 6 levels of service has an astronomically large number of (6^{100}) possible configurations. In that case, does such a system show any organization at all? In other words, if there is organization then there is a subset of configurations that is more stable than others. So what is an indicator of such a phenomenon?

Consider a system characterized by a large number of configurations. If we have little or no information about each of these configurations then the only meaningful thing we can express about these configurations is that each configuration occurs with probability ‘p_i’ where ‘i’ ranges over all possible configurations and that all such probabilities are equal. The amount of information I (Applebaum, 1996; Suhir, 1997) that can be obtained from this system in configuration ‘i’ is given by:

$$I = -\log_2(p_i) \tag{12}$$

The expression for information 'I' in equation (12) can be converted from logarithm of base 2 to logarithm of base 10 as follows:

$$\log_{10}(I) = \frac{\log_2(I)}{\log_2(10)} = 3.3219 \times \log_2(I) \equiv \Theta \log_2(I) \qquad (12A)$$

Since Θ is a constant and it will be ignored in the following discussion as the log operation is assumed to be to the base 10.

In the case of a traffic flow network with 'N' segments and 'L' LOS, let us define a fitness function 'V'. The fitness function computes fitness value M_g (equation 6) for each configuration, where 'g' ranges over $[1, L^N]$. Note that 'V' is a one-to-many function, i.e. many configurations have the same value of fitness. Let R_j,(where j <= g), denote the number of configurations with the same fitness values. Since, the values that R_j can take a priori are unknown, R_j can be expressed as a random variable.

$$R_j = \#V(M_g) \qquad (13)$$

where 'j' ranges over an interval [1, r] and as before 'g' ranges over interval $[1, L^N]$. Let P_j denote the probability distribution associated with R_j. Then, the expectation '*E*' of such a distribution is defined as the amount of uncertainty or information entropy 'S' in the system and is given by:

$$S = -\sum_{j=1}^{r} P_j \times \log(P_j). \qquad (14)$$

As was said above, if there is little or no information about the configurations and hence the associated fitness values, all that can be said about such a system is that R_j, i.e. the number of configurations with similar values of fitness occur with equal probability or R_j has uniform probability distribution. In that case such a system is said to have maximum uncertainty or maximum entropy S_{max} (Jaynes, 1979). Such a system is characterized by disorder or disorganization. On the other hand, a decrease in entropy of the system indicates increasing order or organization in the system. So, how does increasing order or decrease in disorder of a system occur?

Entropy 'S' of a traffic network changes according to probability distributions of a random variable 'R', which is determined by the fitness '*f*' or each configuration which in turn depend on the type of flows on various segments. Hence, at any instant, the difference in 'S_{max}' and 'S' is an indicator of the organization or degree of order in the network. Thus we define an order parameter as follows:

$$O = \kappa(S_{max} - S), \qquad (15)$$

where, ‘κ’ is a proportionality constant and ‘O’ represents the inherent organizing capacity of the network.

The above discussion of entropy and order follows from the treatment of uncertainty in information theory. An information theoretic definition of entropy ‘S’ of a system is a measure of uncertainty in a system (Applebaum, 1996; Levine and Tribus, 1979; Wilson, 1969a, 1973a; Haynes, Phillips, Mohrfeld, 1980; Haynes and Phillips, 1981; Haynes and Storbeck, 1978). If we can reduce uncertainty in the information content of a system then we will have reduced the entropy of the system and accordingly its internal fitness will have improved (see equation 15), an indication of organized traffic flow. Let us consider a tiny network of N = 21 segments and L = 6 LOS; this network has a staggeringly high number, 6^{21} configurations. With no prior knowledge of flow conditions, one must assume that each of these configurations is equi-probable. But as soon as we gather traffic flow information on even a small number of segments, the total number of possible configurations decreases. Further, by defining a discrete random variable that takes on values according to a fitness criteria (see equation 13), one can find the probability distribution of such a random variable. From this the computation of entropy and order can be carried out as suggested in equations (12) through (15). If one is able to maintain a specific level of service on these segments, then every time we make an observation we are certain to find a specific level of service. Then the probability of such segments is 1 and the information content is zero, since log(1) = 0. As flow of traffic on a definite number of segments becomes certain, the overall entropy of the network decreases, indicating an increase in organization of the network. The surveillance equipment to monitor traffic flows (Fig. 2) on segments of a network can provide information that would reduce uncertainty in traffic flows and increase its fitness. Additionally, if TMC is able to maintain higher flows on different segments, it will modify the probability distribution of the flows. This in turn will modify fitness landscapes that are favorable for efficient flows.

Alternately, measurement of flows on all segments gives us the current state of the network. In terms of the traffic flow landscape we now know the vertex that represents the state of the network. If this vertex corresponds to a good fitness then TMC can maintain flows in the network corresponding to that fitness level. On the other hand if a traffic network is on a bad fitness vertex, then the TMC can take measures to improve the fitness of the network and possibly evolve towards a region of better fitness on the fitness landscape. The information on network flows could be used as input for ITS (Intelligent Traffic System) technologies such as ATMS (Advanced Traffic Management System) and ATIS (Advanced Traveller Information Service). In theory TMCs can be distributed across a traffic flow network, each TMC monitoring and managing a subset of segments of the network and helping to maintain efficient network flows. Note that we do not address issues of interfacing all these TMC, in fact, each TMC is considered to be in operation independent of all the others. How to co-ordinate TMCs will be addressed in reports on future research. One of the features of the fitness landscape is its relative independence of factors such as the fitness values and variations in parameter ‘L’ (Kauffman, 1993). Since the traffic landscapes are mainly dependent on values of ‘N’ and ‘K’, each TMC should be able to develop processes to maintain a level of service on segments of a network that would always give a better fitness configuration.

7. Conclusion and Future Directions

Defining fitness vectors for traffic flows on a road network creates a rugged fitness landscape. For large values of N (the number of segments in a network) with K segments ($K < N$) influencing each of N segments, a very complex traffic flow fitness landscape is generated. For $K = 0$, a simple traffic flow landscape with a single global optimum is obtained. Such a traffic network has all it’s segments independent of each other. As the value of K increases, more complex traffic flow landscapes evolve. The other extreme occurs when $K = N - 1$. In this case each segment influences the flow of traffic on all other segments. This type of landscape has an infinite number of local optima and is very rugged. A fitness landscape with $K = N - 1$ indicates a network in which an incident on any one of the segments affects flows across the entire network. It is a highly unstable network. On the other hand, a network with $K = 0$, an incident affects only local flows. The traffic flow landscape model depends primarily on the values of ‘N’ and ‘K’. A TMC like system can influence flows on different segments of a part of a network such that over-all traffic flows across that part of the network can be improved by reducing the level of uncertainty (entropy) in the flows. In other words, increasing organization or order in the network. Future work would involve:

1. Treating variations in flow levels as akin to the concept of mutation (alleles) from evolutionary biology (see Appendix B for brief description of Kauffman’s NK model). In general, flow levels in an urban traffic network are characterized by LOS of C and D, i.e., moderately congested but show variations represented by LOS of A, B (free flow and near free flow), E (congested) or F (highly congested). Most of the mutations in biology are deleterious, i.e. they reduce species fitness. Similarly, traffic flows of LOS of E and F are more frequent and characterize congested traffic network rather than the desirable variation of LOS of A and B type.
2. Interfacing of a network of TMCs across a network.
3. Determination of ranking of segments in a network according to the fitness criteria.
4. Determination of response times to adjust to traffic incidents.

5. Carrying out impact analysis for different scenarios of traffic flows.
6. Exploring the possibility of integrating the model with other dynamic traffic management models based on ITS technology.

Acknowledgment

The authors wish to express their sincere thanks to Dr. J. H. P. Paelinck for his valuable suggestions. NSF/ EPA Grant #SES-9976483 “Social Vulnerability Analysis; NSF Grant #ECS-0085981 “Road Transportation as a Complex Adaptive System and the School of Public Policy’s Center for ITS Evaluation and Implementation. The authors express their appreciation for support from DOT #DTRS98-G-0013.

References

Applebaum D. (1996). Probability and Information: an Integrated Approach. Cambridge University Press. Cambridge, UK.

Arnott R. and K. Small (September-October 1994). “The Economics of Traffic Congestion,” *American Scientist*, **82**, 446–455.

Coleman S., J.B. Hartle, T. Piran and S. Weinberg. (1991). Quantum Cosmology and Baby Universes. Singapore: World Scientific.

Brian A.W. (1989). “Competing Technologies, Increasing Returns, and Lock-in by Historical Events,” *Econ. J.*, **99**, 116–131.

Fabian A.C. (1998). Evolution: Society, Science and the Universe. Cambridge University Press. Darwin College, Cambridge. U.K.

Friesz T.L., D.H. Bernstein and R.R. Stough (1996). “Dynamic Systems, Variational Inequalities and Control Theoretic Models for Predicting Time-Varying Urban Network Flows,” *Transp. Science*, **30**, 1.

Friesz T.L., D.H. Bernstein, N.J. Mehta, R.L. Tobin and S. Ganjalizadeh, (1994). “Day-to-day Dynamic Network Disequilibrium and Idealized Driver Information Systems,” *Operations Research*, **42** (6), 1120–1131

Friesz T.L., Bernstein D.H., R.L. Tobin and B.W. Wie, (1993). “A Variational Inequality Formulation of the Dynamic Network User Equilibrium Problem,” *Operations Research,* **37** (6), 893–901.

Haynes K.E. and J. Storbeck (1978). “The Entropy Paradox and the Distribution of Urban Population,” *Socio-Economic Planning Science*, **12**, 1–6.

Haynes K.E., F.Y. Phillips and J.W. Mohrfeld (1980). “The Entropies: Some Roots of Ambiguity,” *Socio-Economic Planning Sciences*, **14**, 137–145.

Haynes K.E. and F.Y. Phillips (1981). “Minimum Information Models and Individual Choice: Application to Linear and Loglinear Regression and Competitive Spatial Interaction Models,” Ed. G.E. Lasker, *Applied Systems and Cybernetics*, **II**, 733–740.

Herman R. (1982). “Remarks on Traffic Flow Theories and the Characterization of Traffic in Cities.” In *Self Organization and Dissipative Structures*, W.C. Schieve and Allen P.M. (eds.). University of Texas Press, Austin, Texas, 260–284.

Holland J.H. (1995). “Hidden Order: How Adaptation Builds Complexity,” Addison-Welsley, Reading, MA.

Horowitz E. and S. Sahni (1985). "Fundamentals of Computer Algorithms," Computer Science Press Inc, Rockville, MD, 20850.

Jaynes E.T. (1979). "Where Do We Stand in Maximum Entropy?" Eds. R.D. Levine and M. Tribus, *The Maximum Entropy Formalism*, The MIT Press, Cambridge, 15–118.

Kauffman S.A. (1993). *Origins of Order*, Oxford University Press, New York.

Kauffman S.A. and S. Levin, (1987). "Towards a General Theory of Adaptive Walks on Random Landscapes," *J. Theoretical Biology*, **128**, 11.

Koutsopoulos H.N. (1995). "Traffic Simulation Models: Overview and Basic Elements," MIT Summer Professional Program 1.10s, *Modeling and Simulation for Dynamic Transportation Management Systems*.

Krugman P. (1994, May). Complex Landscapes in Economic Geography. *The American Economic Review*, **84** (2), 412–417.

Krugman P. (1995). "Development, Geography and Economic Theory," MIT Press, Cambridge. MA.

Lave C.A. (1985, December). "Speeding, Coordination, and the 55 MPH Limit," *American Economic Review*, 1159–1164.

Linde A. (Nov. 1994). The Self-Reproducing Inflationary Universe. Scientific American, p. 35.

Mahmassani H. (1995). "Dynamic Traffic Assignment for ATIS/ATMS Operation and Evaluation: Issues, Formulations," Algorithms, Implementation and Numerical Experiments, *MIT Summer Professional Program 1.10s, Modeling and Simulation for Dynamic Transportation Management Systems.*

Mahmassani H. and R. Herman (1987). "Interaction of Trip Decisions and Traffic Systems Dynamics," *European Journal Operations Research* **30**, 304–317.

Mahmassani H. and R. Herman (1990). "Interactive Experiments for the Study of Tripmaker Behavior Dynamics in Congested Commuting Systems," In *Developments in Dynamic and Activity-Based Approaches to Travel Analysis*, P. Jones (ed.). Avesbury, Gower Publishing, Aldershot, England, 272–298.

Mahmassani H.R., Jayakrishnan and R. Herman (1990). "Network Traffic Flow Theory: Microscopic Simulation Experiments on Supercomputers," *Transaction Research A* **24A**, 149–162.

Mahamassani H.S., T. Hu and R Jayakrishnan, (1992). "Dynamic Traffic Assignment and Simulation for Advanced Network Informatics," *Proceedings of the Second International Capri Seminar on Urban Traffic Networks*.

Mahmassani H. and S. Peeta (1992). "Dynamic Assignment for Route Guidance Decisions," presented at the *39th North American Meetings of the Regional Science Association*.

Murchland J.D. "Braess' Paradox of Traffic Flow," *Transp. Research* **4** (1970), pp. 391–394.

Paelinck J.H.P. and L.H. Klaassen (1979). "Spatial Economics," Saxon House, Teakfield Ltd., Westmead, Farnborough, Hants., England.

Suhir E. (1997). Applied Probability for Engineers and Scientists. McGraw-Hill, New York.

Tanner J.C. (1961). "Factors Affecting the Amount of Travel," *Road Research Laboratory*, Technical Paper # 51, HMSO, London.

TRANSIMS (1995). Sandia National Laboratories, Santa Fe, NM, USA.

TIGER/Line™ Files (1991). Census Feature Class Codes, Appendix E, TIGER/Line Census Files, 1990 prepared by the Bureau of the Census, Wash. D.C., The Bureau.

TRB (1985). Special Report No. 209, Highway Capacity Manual, National Research Council, Wash., D.C.

Tu Pierre N.V. (1992). Dynamical Systems: an Introduction with Applications in Economics and Biology. Springer Verlag. Berlin, New York.

Wilson A.G. (1969a). "The Use of Entropy maximizing models in the theory of trip distribution, mode split and route split," *Journal of Transp. Econ. Policy*, **3**, 108–126.

Wilson A.G. (1973a). “Further Developments of Entropy Maximizing Transport Models,” *Transportation Planning Technology*, **1**, 283–306.
Wilson E.O. (1995). On Human Nature. Penguin, London.
Wilson R. (1988). “An Intro. to Dynamic Data Structures,” McGraw-Hill Books (UK) Ltd., England.

Appendix A

Consider a road segment of length l and width w. Then the total area A of the road segment is given by:

$$A = L^D * W^D \text{or if we express w as fraction of l then}$$
$$A = L^D * (\gamma * L^D)\text{or alternately it may be expressed as follows:}$$
$$A = \gamma * L^{2D}, (A1) \tag{A1}$$

Equation (A1) can be re-written as:

$$A \alpha L^{2D} \tag{A2}$$

where $D = 1$, the dimension of the road segment. Now consider a stream of vehicles traveling on the segment of the road. Thus at any instant there are finite number of vehicles occupy a finite amount of space on a section of the road.

Since, the vehicles on a road are discrete objects and occupy finite and discrete amount of space, we can express the total area occupied by the vehicles as follows:

$$a = n * l^d * w^d = n * l^d * (\delta * l^d), \tag{A3}$$

where a is the average area occupied by a vehicle of average length l and average width w and δ is a fractional measure for converting w into l. Equation (A3) may be expressed as:

$$a = n * \delta * l^{2d} \quad \text{or} \quad a\alpha(n * l^{2d}) \tag{A4}$$

Let us express the average value of vehicle length l in terms of the length of section of the road, then equation (A4) can be written as:

$$a = n * \delta * l^{2d} \quad \text{or} \quad a\alpha(n * (\varepsilon * L)^{2d}) \tag{A5}$$

From equations (A2) and (A4) the density of vehicles occupancy ρ may be expressed as:

$$\rho \alpha \frac{n * (\varepsilon * L)^{2d}}{L^{2D}} \quad \text{or} \quad \rho \alpha \frac{n * \varepsilon^{2d} * L^{2D}}{L^{2D}}. \tag{A6}$$

We can express the density function ρ by introducing a proportionality constant β in equation (A6) and get the following equation:

$$\rho = \frac{\beta * n * \varepsilon^{2d} * L^{2d}}{L^{2D}} = (\text{const.}) * L^{2(d-D)}. \tag{A7}$$

Taking logarithm on both sides of equation (A7) gives us the following equation:

$$\log(\rho) = \log(\text{const.}) + 2(d - D) * \log L. \tag{A8}$$

Since, $D = 1$ we can get the following equation:

$$d = \frac{\log(\rho) - \log(\text{const.}) + 2\log(L)}{2 * \log(L)}, \tag{A9}$$

From equation (A8) we can get an expression for d as follows

$$d = 1 + \frac{\log(\rho) - \log(\text{const.})}{2 * \log(L)} \tag{A10}$$

Value of d varies between a minimum of zero (free flow) and maximum = 1 (blocked segment) (see equation A10). This computed value of d can be used as a measure of the level of service for assigning fitness values to sections of roads.

Appendix B

Kauffman's NK Model

The NK model describes emergence of order in biological systems as a result of a slew of complex, random, epistatic (non-reciprocating and inhibitory) binary interactions among the most fundamental agents of self-organization, the genes. A population of genes (genotypes) evolves over a fitness landscape (a type of hill-climbing) as it adapts to changes in the environment. To get a better understanding of the NK model, given below are definitions of the biological terms. Gene is the basic unit of inheritance. Genotype is a possible configuration or arrangement of genes. Allele is a variation of a gene. Fitness is any “well defined property” and the fitness landscape is a distribution of this property across an ensemble (Kauffman, 1993).

The NK model of the evolutionary biologist Kauffman explains how a variety of genotypes is able to adapt to so-called rugged fitness landscapes of the environment in which these genotypes evolve. The ‘N’ stands for number of genes and ‘K’ stands for number of interactions any single gene has with other genes. Each gene may have ‘L’ alleles. Alleles are the variations in each gene that give rise to a physical trait such as eye color of blue, brown, black etc. Each gene contributes to the overall fitness of a genotype. At the same time each is influenced by ‘K’ genes that are either nearest neighbors or are spatially separated from the gene. Thus the result of all the interactions between ‘N’ genes and ‘K’ influencing genes is a fitness landscape with multiple peaks and valleys. The peaks are associated with fitness values. Depending on the value of ‘K’, the landscape varies from a simple profile (K = 0) to one with a very complex profile (K = N – 1). The former (K = 0) refers to an environment in which each gene is independent of all its neighbors and the latter refers to a situation when each gene is influenced by all the genes (K = N – 1) in a genotype.

Chapter 7

Network Connectivity Models: An Overview and Empirical Applications

Aura Reggiani[1] and Sandra Vinciguerra[2]

[1]Department of Economics, Faculty of Statistics, University of Bologna, Piazza Scaravilli, 2, 40126 Bologna, Italy, E-mail: aura.reggiani@unibo.it

[2] Department of Human Geography and Urban and Regional Planning, Faculty of Geosciences, Utrecht University, Heidelberglaan 2, P.O. Box 80115, 3508 TC Utrecht, The Netherlands, E-mail: S.Vinciguerra@geo.uu.nl.

Abstract In recent years great attention has been paid to complex networks and to their related theories and models.

In this context, the concepts of Small World and Scale Free networks come to the fore. A Small World network (SW) is based on 'six degrees of separation', or the notion that everyone in the world is related to everyone else through at most six acquaintances. Small World networks are similar in that they have a high degree of local clustering or cliquishness and a relatively short average minimum path, like a completely random network. A Scale Free (SF) network is principally characterized by an abundance of nodes with only a few links, while a very small number of nodes have a very large number of links, which are called hubs.

This paper examines the above issues in the context of regional dynamics from both the methodological and empirical view point.

In particular, the first part aims to provide an overview on the concepts, insights and research perspectives in spatial economics concerning SW and SF networks, in the light of their physical and statistical characteristics, e.g. diameter, clustering coefficient and vertex connectivity degree distribution.

The second part of the paper attempts to explore some empirical applications in order to point out common features that characterize these networks in Socio and Spatial Economic Sector.

The paper will end with methodological observations concerning the role of exponential/power law in the spatial economic literature.

Keywords: spatial economics; network science; power laws; Scale Free; Small World

1. Introduction

In the last few years literature on complex networks, and especially on "Small World" (SW) and "Scale-Free" (SF) networks, has increased greatly. Researchers in many fields can easily find applications on these models, for example in social science, biology, economics, technology and telecommunication (see, e.g. Gorman, 2005; Patuelli et al., 2006; Schintler et al., 2005a, 2005b).

Recently, attention has also been directed to the study of the connectivity properties and the topology of complex systems, especially on "who is connected to whom". In this context, complex systems are often modelled by graphs, composed by vertices or nodes (representing the elements of the graph) and edges or links (representing the interactions or connections between the single elements of the graph). Graph theory has been the conceptual framework of the recently developed network models, such as SW and SF networks.

In the next section, Random, SW and SF networks will be concisely illustrated from a theoretical-methodological view point. Section 3 will present some methodological remarks concerning SW and SF models, while Section 4 will be devoted to outline their applications in the space-economy. Section 5 will conclude the present chapter by highlighting methodological observations concerning the role of SF models in the real spatial economic networks.

2. Random, Small World and Scale Free Networks: a Concise Overview

The Random Network (RN) – introduced in graph theory by Erdős and Rényi (1959) – identifies networks whose vertices are randomly connected to each other. A RN is usually displayed as a graph composed by a number of nodes N randomly connected by links with probability p. The cumulative distribution[1] of nodes degree (*connectivity degree*) follows a Poisson distribution, which means that the majority of the nodes on the network have the same number of links, nearby the average degree $<k>$; nodes that deviate from this average are rare (see Figure A1 in the Annex). A RN can be compared to a homogeneous system which gives accessibility to the majority of the nodes in the same way.

However, clustering characters often emerge in dynamic networks, by showing that economies of density seem to run parallel to economics of motion. For example, thinking to the nodes characterizing social and telecommunication

[1] The cumulative distribution function describes the probability distribution of a real-valued random variable X, $F(x) = P(X \leq x)$.

networks, as well as to electricity cables and transport networks, it is clear that possibly not all have the same number of links, and that these links are not randomly placed. In these networks, clustering features seem more suitable. They can be detected by the SW network.

The SW phenomenon[2] has been formalized by Watts and Strogatz (1998) in their study on topological properties of a network. These authors developed a "SW model" focusing their attention on two different network parameters, the *diameter*[3] and the *clustering coefficient*.[4] In particular, they demonstrated that a SW network is identified by a high clustering coefficient C and a short diameter L while a RN is characterized by short diameter and low clustering coefficient (see Table 1 and Figure A1 in the Annex).

Attention must be paid to the difference between "SW effect" and Watts and Strogatz' "SW network": the first can be seen as a property of several network models and identifies only the fact that the diameter of a network is short enough to reach in a few steps every vertex of the network. The "SW network" describes something more: in addition to the short diameter, the model is characterized by a high clustering coefficient. Referring to the "SW effect", there is it possible to claim that even RNs present a SW property, in the sense that they have a short diameter. However, a RN does not belong to SW networks because it lacks in high clustering coefficient. It should be noted that 'SW networks' has been shown to emerge in several systems, from neural networks (Watts and Strogatz, 1998) to the World Wide Web (Adamic, 1999), to the power grid of the Western United States (Watts, 1999), to the diffusion dynamics of infectious diseases (Boguñá et al., 2002).

Close to the SW network a further network approach, the so-called Scale Free (SF) – which came to the fore in the '90s – it is worth to be investigated. The SF network – conceived of by Barabási and Albert (1999) – is based on two mechanisms:

- *Incremental Growth*: the networks are dynamic systems; the number of nodes is not static; the network grows with time;

---

[2] Taking inspiration from the "six degrees of separation" experiment on the US social network made by Milgram (1968).

[3] Diameter: (also called 'characteristic path length') measures the typical separation between two generic vertices within the graph (a global property) in terms of number of links composing the path. The formula is the following one: $L(G) = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}$, where d_{ij} is the number of links of the shortest path to reach vertex j from vertex i, and N is the number of nodes in the network.

[4] Clustering coefficient: measures the cliquishness of a typical neighbourhood (a local property). The formula is the following one: $C(G) = \frac{1}{N} \sum_i C_i$, where C_i is the ratio between the number of existing links and the maximum number of possible links in G_i, sub-graph of the neighbourhood of vertex i, N is the number of nodes of the network.

- *Preferential Attachment*: the new vertices are not randomly connected to the existing nodes; they are linked with greater likelihood to high connectivity degree vertices.

According to Albert et al. (1999), the implementation of these two mechanisms in a model is sufficient for the generation of SF networks.

With regard to its *physical measurement*, a SF network presents a diameter shorter than a SW. In addition, a SF network presents a very high clustering coefficient which decreases with the increase of the number of vertices in the network (Table 1). Thus we can underline that a SF network strongly embeds the SW properties.

Concerning its *statistical measurement*, a SF network is characterized by a power law vertex connectivity degree distribution (Table 1). It should be noted that the power law/SF property is not frequently found in the data distribution. On the contrary, the Poisson or exponential law – which can be methodologically connected to a RN – seems to characterize most of the distributions. An appropriate way of detecting the type of network distribution, is the vertex connectivity degree distribution. This is defined as the probability $P(k)$ of finding nodes with k links: $P(k) = N(k)/N$, where $N(k)$ is the number of nodes with k links, and N is the total number of nodes within the network (Table 1).

The name "Scale Free" is originating from the fact that the power law distribution does not change its form no matter which scale is used to observe it. In other words, even though it is possible to change the scale of the distribution data, the form of the distribution remains the same.

Table 1 shows that, concerning the form of the vertex connectivity degree distribution in a network, we can recognize the following main typologies:

- A Poisson distribution,[5] which identifies a RN[6] (Erdős and Rényi, 1959)
- A power law distribution,[7] which identifies a SF network (Barabási and Bonabeau, 2003).

A power law distribution is different from the exponential/Poisson distribution. It refers to networks displaying: a) an abundance of nodes with just a few links; b) a small number of nodes with a very large number of links (the so-called 'hubs'[8]) (see Figure A3 in the Annex). It should be noted that the

[5] Poisson distribution: $P(k) \propto e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$.

[6] "*RNs are also called exponential, because the probability that a node is connected to k other sites decreases exponentially for large k*" (Barabási and Bonabeau, 2003, p. 52).

[7] Power law distribution: $P(k) \propto k^{-\gamma}$.

[8] Hub: a single vertex with a large number of connections, the so called 'preferential node' (Barabási and Oltvai, 2004).

Table 1. Overview on characteristics of RN, SW and SF networks. N indicates the number of vertices of the network, k indicates the variable "number of links connected to a node" and <*k*> is the average degree of the network, and γ is the exponent degree of the power law distribution.

		RN network	SW network	SF network
Physical Measures	**Diameter L**	Short	Short scales as $L \sim \ln N$	Very short scales as $L \sim \ln \ln N$
	Clustering coefficient C	Low	High	High, but it decreases with the increasing of the network size N
Statistical Measures	**Vertex connectivity degree distribution**	Poisson $P(k) \propto e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$	Similar to the RN, decaying exponentially for a large set of vertices	Power law $P(k) \propto k^{-\gamma}$
	Exponent degree			$2 < \gamma < 3$

different forms of the vertex connectivity degree distributions mirror different economic meanings. The RN can be considered as a 'democratic model' (i.e. most vertices have approximately the same number of links, close to the average degree), where "*vertices with large connectivity are practically absent*" (Barabási and Albert, 1999, p. 510). The SW/SF network – owing to its clustering character – can be considered 'hierarchical' in its economic functions and activities.

In this context, the value of the coefficient in each type of statistical distribution plays a critical role, since it determines the shape of the function.

Concerning the value of the exponent degree γ of the power law distribution, we can identify the following typologies:

a) in the case that the exponent degree is $\gamma = 2$, a hub-and-spoke[9] network emerges;
b) if the exponent degree varies between 2 and 3, the SF network presents a much smaller diameter than the diameter of a RN or a SW network (Cohen and Havlin, 2003). In general, the diameter of a SW network scales as $L \sim \ln N$. However for this particular range of values for the exponent degree $(2 < \gamma < 3)$, the diameter of a SF network scales as $L \sim \ln \ln N$, where N is the number of vertices (Table 1). Thus a SF network can be considered as

[9] Hub-and-spoke: is a topology of network that refers to the use of a central node to coordinate activities between the other nodes, which are not connected between them but all to the central one; like a bicycle wheel, a location is selected to be a hub and the paths that lead from points of origin and destination are considered spokes.

"Ultra Small World" network: "*Although the small-world effect is considered as a property of random networks, scale free networks are ultra small*" (Barabási and Oltvai, 2004, p. 106). In addition a hierarchy of hubs emerges;

c) in the case of an exponent degree equal to 3, the power law function decays faster than in the various cases;

d) in the case of an exponent degree $\gamma > 3$, hubs are no longer important and the diameter scales again as for a SW network. Moreover, according to Barabási and Oltvai[10] (2004), the SF network's behavior becomes similar to that of a RN. In particular, by increasing the coefficient value even more, the number of hubs in the network decreases, while all the other nodes have only a few links, until there are no more hubs and the network's behaviour reverts to being similar to a RN, even though the degree distribution does not turn into an exponential function.

A final observation concerns the 'history' of power law in science. Power law distribution is not something new in science. It has already been observed in nature that many real systems fit a power law function (see, Chapter 1 in Bak, 1996). It is often associated with the theory of the Rank Size Rule[11] (Zipf, 1949), which describes the relationship between the ranks of cities and their population (see Figure 1). It should be noted that Nitsch (2005) determines that the distributions of city population are typically best described by a power law with exponent 1.0.

We can then assert that the conventional Rank Size/power law, conceived by Zipf, has been in a way revisited – in a network perspective – by Barabási and his collaborators. The novelty in the SF network is represented by the clustering concept, which also offers a socio-economic aspect to the empirical evidence of Zipf's power-law.[12]

After this outline on RN, SW and SF networks, in the next section we will focus on some of the methodological remarks, mainly concerning the differences between SW and SF networks.

---

[10] Barabási and Oltvai (2004) affirm that a SF network emerges only for an exponent value of the power law distribution between 2 and 3, but they do not clearly explain the reason.

[11] Rank Size Rule's (Zipf's law) formula is the following: $P_n = P_1/n$, where P_n is the population of town ranked n, P_1 is the population of the largest town and n is the rank of the town.

[12] The *rank/frequency plot* – introduced by Zipf in 1949 – of the cumulative distribution function $P(x)$ is defined, for example, as the occurrence frequency of words in a text, or the probability to find the same word (e.g. "the") in a text. $P(x)$ is the fraction of words with frequency greater than or equal to x. Alternatively, it is possible to plot the number of words (absolute value) with frequency greater than or equal to x. The difference between the two methods is only in the normalization of the data. Then the values associated to the words will be sorted into the decreasing order of frequency and plotted with their ranks as a function of their frequency (Newman, 2005).

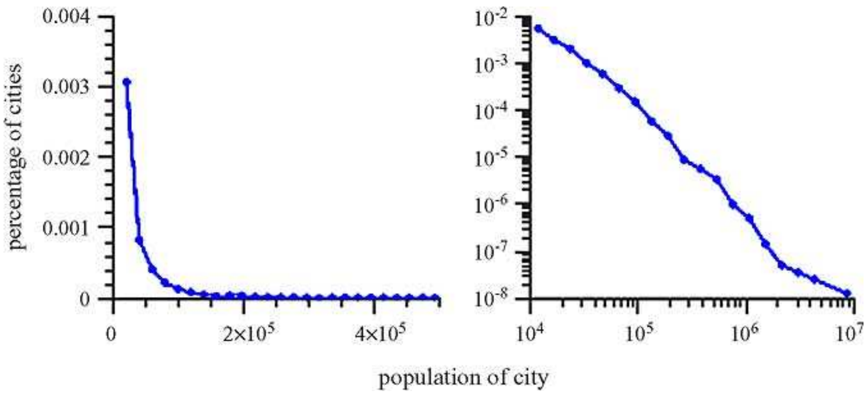


Figure 1. Example of an application of the Rank Size Rule on the population of all US cities where the population is 10000 or more, in year 2000. Both histograms plot the same data, but the one on the right is plotted on a double logarithmic scale (Newman, 2005, p. 324).

3. Scale Free and Small World Networks: Methodological Reflections

Several studies were developed in recent years in order to better investigate from a theoretical, methodological and empirical view point if these type of networks emerge in real complex systems.

The question is how to define a SF network. Amaral et al. (2000, p. 11149) assert that SF networks are a sub class of SW networks. They distinguish three classes of SW networks:

1. Scale-free networks, characterized by a vertex connectivity degree distribution that decays as a power law;
2. Broad-scale networks, characterized by a vertex connectivity degree distribution that has a power law regime followed by a sharp cut-off;
3. Single-scale networks, characterized by a vertex connectivity degree distribution with a fast decaying tail.

Many criticisms have been made on the SF theory, asserting that the SF is too simplistic for the Internet and its definition is often ambiguous (Chen et al., 2002). The introduction of the rewiring principle[13] in the SF model, after incremental growth and preferential attachment by Albert and Barabási (2000), was one of responses to these criticisms.

In addition, in 2002 Albert and Barabási pointed out that SW and SF networks are different types of network phenomena: the former explaining

[13] Rewiring principle: see Note 3.

clustering and the latter explaining the power law for the vertex connectivity degree distribution.

A further interesting definition has been offered by Dorogovtsev and Mendes (2003), who make a distinction between *equilibrium networks* (non-growing networks) and *non-equilibrium networks* (growing networks), by studying connectivity degree distributions (how large their tails are) and the related simple properties. In this context, they classify SW as *equilibrium* networks and SF as *non-equilibrium* networks.

A clear delineation of where SW and SF networks diverge is still missing in the literature, even though scientific efforts in this direction come recently to the fore. For example, Gorman and Kulkarni (2004, p. 9) outline the subsequent SW and SF properties: *"It can be safely said that the two models are inter-related and that generally speaking, SF networks exhibit the clustering and short average path length of SW networks, but not all SW networks exhibit the power law distribution of SF networks"*.

It should be noted that SF networks might also be denominated "Scale Rich" because of the co-existence in the networks of nodes of widely different degrees (scales), ranging from nodes with one or two links connected to major hubs, to nodes with a high number of links (Barabási and Oltvai, 2004, p. 104).

In the next section we will present a brief review of empirical applications of SF networks mainly carried out in the socio-economic-spatial field, in order to point out their common features.

4. Empirical Applications to Spatial Economic Networks

4.1 A Concise Overview

On the basis of the previous methodological observations, a methodological concern regards the identification of a SF network versus a SW/RN. Clearly, the empirical evidence – from the statistical viewpoint – of a power law versus an exponential law might be considered as a fundamental procedure in this respect.

A first schematic approach is to take empirical data on network connectivity and plot them in order to investigate, with statistical tests, if these data fit a power law function. In this context, it is interesting to observe the differences between power law and exponential function.

The exponential law is a function frequently utilized in literature, and in the network case it should represent the probability of node connection in a RN. It essentially shows a decay function that reaches zero for high values of the x-axis. On the contrary, the power law maintains a higher tail and never reaches zero, even in the long run. Figure 2 shows an example of the two types of function, by considering an exponent degree equal to 1. In particular, in Figure 2 the distance between the two curves decreases in the long run. In

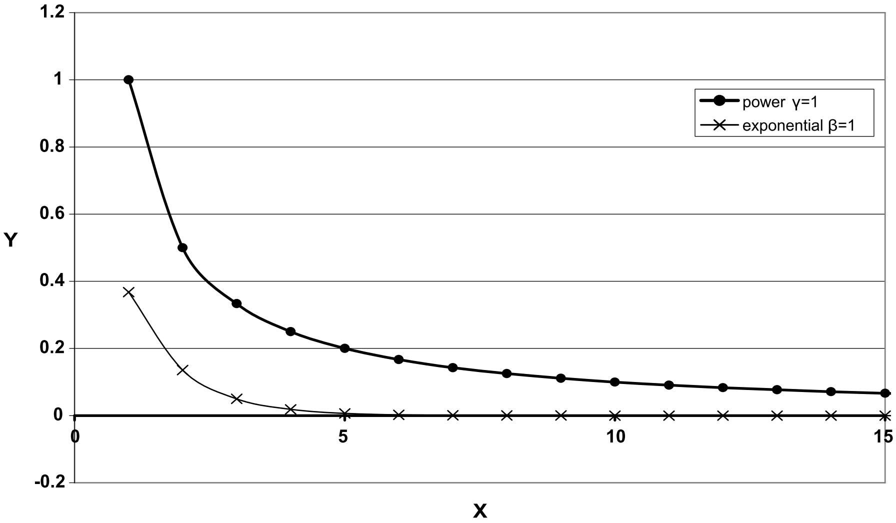


Figure 2. Power law vs. exponential law functions: simulation experiments for an exponent degree value equal to 1.

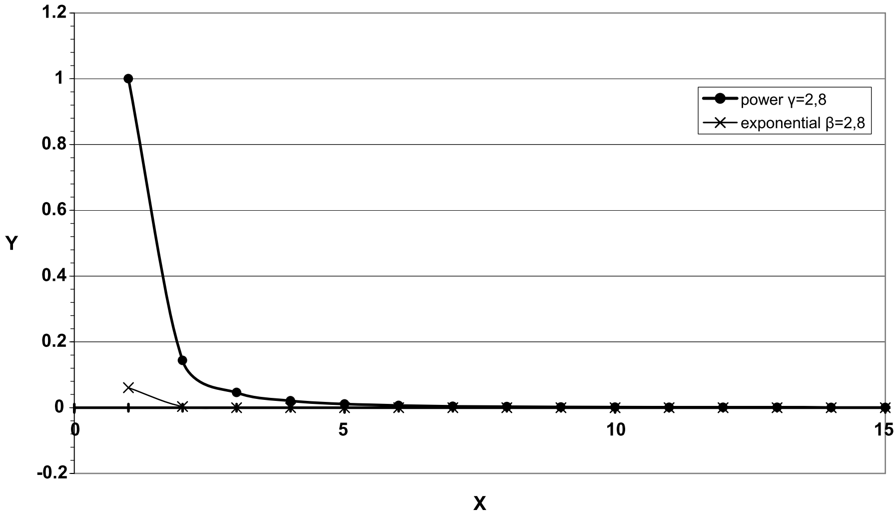


Figure 3. Power law vs. exponential law functions: simulation experiments for an exponent degree value equal to 2.8.

addition, this 'vertical' distance between the two curves is even smaller by increasing the value of the coefficient (see Figures 3 and 4). The related log-log transformations are illustrated in Figures B1, B2 and B3 in Annex B.

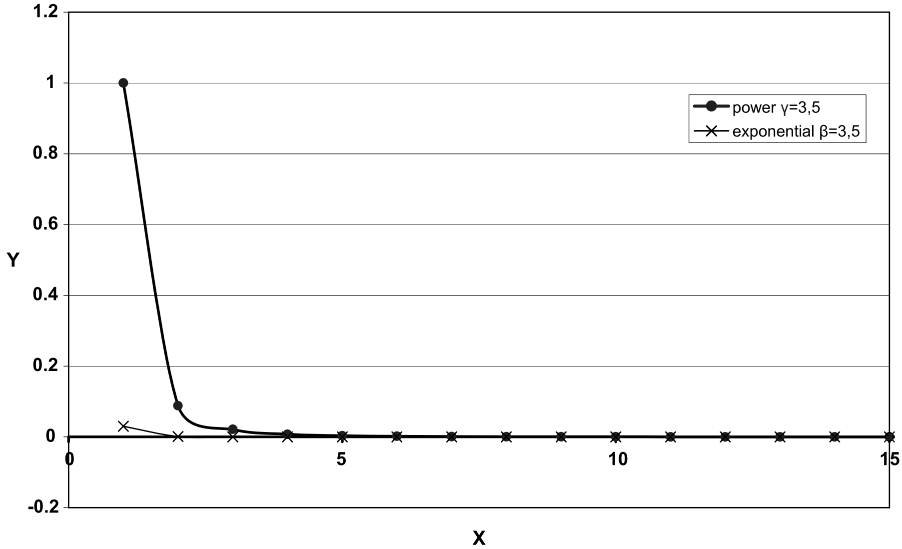


Figure 4. Power law vs. exponential law functions: simulation experiments for an exponent degree value equal to 3.5.

On the basis of the previous considerations, we can extract some common elements in the applications carried out in the spatial economic field. Table 2 – although not exhaustive – shows the most recent SF applications in both the spatial economic and social sector.

First, it can be stated that the majority of the networks presented in Table 2 shows a coefficient degree γ – in the related power law distribution – between 2 and 3. Since, according to Barabási and Oltvai (2004), this range of values identifies the existence of a SF network, we can deduce that the so-called 'virtual' networks, such as internet and market investments, show SF characteristics, while the physical networks, such as commuting and telephone calls, show a power law with a very low exponent degree (and thus the absence of a SF network of Barabási type).

This result confirms the 'irrelevance' of distances in SF networks. Consequently, it seems difficult to detect the emergence of a Barabási SF network in physical systems, like commuting, where geographical distance is relevant. In general, it seems that in the networks where distance becomes important in the preferential attachment mechanism, there is a 'practical' reason that obstructs the spreading of an ideal SF network, i.e., the relative costs in terms of time and money. It should be noted that SF networks are simple models, without the constrictions which inevitably emerge in empirical networks. Thus, not always the empirical networks indicate the presence of SF properties, but at the same time they cannot be classified as RNs because connections are not always

Table 2. Empirical values of the coefficient γ which characterise the degree distribution of power law networks (size = number of nodes; <k> = average degree).

Networks	Size	$< k >$	γ	References
Networks without SF Features			$\gamma < 2$	
Sardinian inter-municipal commuting network	375	43.33	–	De Montis et al., 2005
Internet infrastructures EU	209	–	0.54	Schintler et al., 2005b
Commuting network in Germany	439	18.21	1.11	Russo et al., 2006
Co-authors, SPIRES	56627	173	1.2	Newman, 2001
E-mails	59912	2.88	1.81	Ebel et al., 2002
Internet infrastructures USA (2000)	326	–	1.83	Schintler et al., 2005a
WWW, site	260000	–	1.94	Huberman et al., 1999
SF Networks			$2 < \gamma < 3$	
World-Wide Airport Network	3880	9.7	2	Barrat et al., 2004
Phone-call	53000000	3.16	2.1	Aiello et al., 2000
Co-authors, neuro	209293	11.54	2.1	Barabási et al., 2002
Internet, domain	3015–4389	3.42–3.76	2.1–2.3	Faloutsos, 1999
WWW	40000000	7	2.1–2.38	Kumar et al., 1999
WWW	325729	4.51	2.1–2.45	Albert et al., 1999
WWW	200000000	7.5	2.1–2.72	Broder et al., 2000
Market Investments – NAS	2053	–	2.22	Garlaschelli et al., 2005
Market Investments – NYS	240	–	2.37	Garlaschelli et al., 2005
Internet, router	150000	2.66	2.4	Govindan, 2000
Internet, router	3888	2.57	2.48	Faloutsos, 1999
Co-authors, math	70975	3.9	2.5	Barabási et al., 2002
Words, co-occurence	460902	70.13	2.7	Cancho et al., 2001
Words, synonyms	22311	13.48	2.8	Yook et al., 2002
Market Investments – MIB	3063	–	2.97	Garlaschelli et al., 2005
Networks without SF Features			$\gamma > 3$	
Citation	783339	8.57	3	Redner, 1998
Comic Book Characters	6486	14.9	3.12	Alberich et al., 2002

implemented randomly. Some kind of 'preferential attachment' certainly exists, even though not exactly of the SF type (Russo et al., 2007).

It is moreover important to identify a SF network because of its strong features in terms of robustness and vulnerability. In the case of a random attack on nodes, the SF network will strongly persist, because a random attack will probably damage nodes that have only a few connections, which are the majority; nevertheless in case of an attack against the main hubs, the network will easily be fragmented. On the other hand, RNs are weak against a random attack which will cause the split of the network.

Recent simulations on SF networks demonstrate that the damage of just a few of the major hubs will provoke the crash of the whole system (Gorman et al., 2007). Consequently, it is important to identify hubs in the network in order to prevent targeted attacks and preserve the system (Gorman, 2005). In this context, further studies are needed in order to understand whether connections can be redistributed over time after a node is added or destroyed; undoubtedly, this is strongly related to the nature of the network, especially on which nodes and links it is identified.

Since the network typology seems either to facilitate or hamper the emergence of the SF characteristics, it would also be useful to identify the critical factors leading to such developments. Hence, it might be worthwhile to explore new models to be able to grasp those empirical networks which do not match all the SF network features, for example by introducing the geographical distance as a decision tool in the "preferential attachment" mechanism. Further decision tools strongly influencing the topology of the network might be the cost function of the spreading of the network and/or the cost of adding a new node/link. For example, if the cost of nodes is relatively higher than the related connections (eg., building a new airport in the airport transport network), we would expect the emergence of a densely connected network; on the contrary, if the cost of a new connection is relatively higher than for a new node (as in the case of the network of the internet backbone infrastructure), we would expect a less densely connected network.

Geographical distance is also important when treating firm networks: social network analysis is more often applied to investigate cooperation and interaction between firms within the same sector. It seems that the location of firms within the same district or the same region (geographical proximity) helps the rising of collaboration and interaction between them, by spreading of innovation and knowledge (Boschma, 2005).

4.2 Additional Functions Detecting Scale Free Networks

In the previous sections we outlined how power law – with a degree exponent varying between 2 and 3 – is usually considered the principal function detecting a SF network.

In addition, it is worthwhile to outline another function able to identify a SF network: the power law with exponential cut-off.[14] This function is similar to the power law, however here the tail decays as an exponential function. This function – introduced by Jeong in 2001 – can be more easily adapted to empirical data. Indeed some of the applications included in Table 2, such as the e-mail network (Ebel et al., 2002), are based on this function (see Figure 5).

The exponential cut-off presents a slightly different form from the power law, but it displays the same SF properties (see Section 2). At the same time, the related coefficient degree follows the SF values previously discussed.

[14] Power law with exponential cut-off formula: $P(k) \sim (k + k_0)^{-\gamma} e^{-(k+k_0)/k_c}$, where k_0 and k_c are constants and γ is the coefficient degree of the Power law function. For details see Vinciguerra (2005).

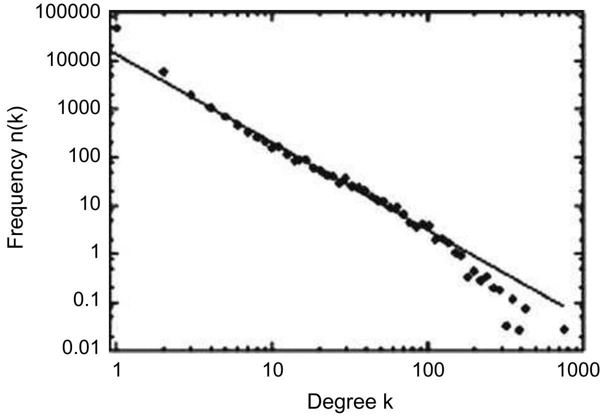


Figure 5. Degree distribution of the e-mail network, on a double logarithmic plot (source: Ebel et al., 2002, p. 035103-1).

5. Concluding Remarks

There is still much to be learnt about complex networks, however some preliminary remarks can be highlighted. First, SF networks, even if dynamic systems, do not consider the "death" of nodes. An example is provided by internet, where every day new web-pages (considered as nodes) are added to the system, while simultaneously other pages "die", and with them all the links to those pages, even if they persist for a long time in the search engines. This means that a lot of links, which are no longer useful, still exist.

Second, the introduction of a new node, as in the case of airport networks, will probably cause a redistribution of connections, not only for the nodes belonging to the same cluster, but also for the whole network. This redistribution is affected: a) by the importance of the node; b) by the nodes in the proximities; and c), by the weight of new connections. The mechanism of growth in the SF model can also cause congestion in a long term period; for example, a hub in an airport can support only a limited number of flights per minute, thus the question is, what will happen when saturation is reached.

Concerning the dynamics of SF networks, no studies – to our knowledge – exist on the persistence of the SF properties over time, in other words, if the network can be identified as a SF from the beginning to the last observation, or if the topology of the network is changing, over time, into a RN.

The problem of forecasting the SF network and its dynamic trajectory is still an open research issue. More experiments, especially in a dynamic framework, are then necessary, in order to understand how the topology of SF networks evolves and whether the hierarchy of hubs is changing over time (e.g. the decay of some hubs for the airport network, or the persistence of Google for the internet).

In conclusion, SF networks appear to be a useful tool in order to detect economic hubs and clusters in spatial economic networks. The main concern is whether a real spatial network can be identified as a SF; hence studies making use of rigorous statistical tests to demonstrate the existence of SF networks are needed. In this context, it is important to explore how cultural, political, economic and technological factors may influence positively or negatively the development of a SF network.

References

Adamic L. A., 1999, "The Small World Web", *Lecture Notes in Computer Science*, 1696, 443–452.

Aiello W., Chung F. and Lu L., 2000, *Proceedings of the 32nd ACM Symposium on the Theory of Computing* (ACM, New York), 171–180.

Alberich R., Miro-Julia J. and Rosselló F., 2002, "Marvel Universe looks almost like a Real Social Network", Los Alamos Archive, cond-mat/0202174

Albert R. and Barabási A. L., 1999, "Emergence of Scaling in Random Networks", *Science*, 286, 509–512.

Albert R. and Barabási A. L., 2000, "Topology of Evolving Networks: Local Events and Universality", *Physical Review Letters*, 85, 24, 5234–5237.

Albert R. and Barabási A. L., 2002, "Statistical Mechanics of Complex Networks", *Reviews of Modern Physics*, 74, 47–97.

Amaral L. A. N., Scala A., Barthélémy M. and Stanley H. E., 2000, "Classes of Small-World Networks", *PNAS* 97, 21, 11149–11152.

Bak P., 1996, *How Nature Works*, Springer-Verlag, Berlin.

Barabási A. L., 2002, *Linked: the New Science of Network*, Cambridge Massachusetts, Perseus Publishing.

Barabási A. L. and Bonabeau E., 2003, "Scale-Free Networks", *Scientific American,* 288, 60–69.

Barabási A. L. and Oltvai Z. N., 2004, "Networks Biology: Understanding the Cell's Functional Organization", *Nature Reviews – Genetics*, 5, 101–113.

Barabási A. L., Jeong H., Ravasz R., Neda Z., Vicsek T. and Schubert A., 2002, "Evolution of the Social Network of Scientific Collaborations", *Physica A*, 311, 590–614.

Barrat A., Barthélemy M., Pastor-Satorras R. and Vespignani A., 2004, "The Architecture of Complex Weighted Networks", *PNAS*, 101, 11, 3747–3752.

Boguñá M., Pastor-Satorras R. and Vespignani A., 2002, "Epidemic Spreading in Complex Networks with Degree Correlations", *Physical Review E*, 66, 047104.

Boschma R. A., 2005, "Proximity and Innovation. A Critical Assestment", *Regional Studies*, 39, 1, 61–74.

Broder A., Kumar R., Maghoul F., Raghavan P., Rajalopagan S., Stata R., Tomkins A. and Wiener J., 2000, "Graph structure in the Web", *Computer Networks*, 33, 309.

Chen Q., Chang H., Govindan R., Jamin S., Shenker S. J. and Willinger W., 2002, "The Origin of Power Laws in Internet Topologies Revisited", *in Proc. of IEEE Infocom.*

Cohen R. and Havlin S., 2003, "Scale-Free Networks are Ultrasmall", *Physical Review Letters*, 90, 5, 058701.

Cowan R., 2004, "Network Models of Innovation and Knowledge Diffusion", MERIT Research Memorandum # RM2004-016.

Crucitti P., Latora V., Marchiori M. and Rapisarda A., 2001, “Complex Systems: Analysis and Models of Real-World Networks”, *Energy and Information Transfer in Biological Systems*, World Scientific Editors.

De Montis A., Barthélémy M., Campagna M., Chessa A. and Vespignani A., 2005, “Emergent Topological and Dynamical Properties of a Real Inter-Municipal Commuting Network: Perspectives for Policy-Making and Planning”, *ERSA – Congress of the Regional Science Association*, Amsterdam.

Dorogovtsev S. N. and Mendes J. F. F., 2003, *Evolution of Networks: From Biological Nets to the Internet and WWW,* Oxford University Press.

Dorogovtsev S. N. and Mendes J. F. F., 2001, “Effect of the Accelerating Growth of Communication Networks on their Structure”, *Physical Review E*, 63, 025101.

Ebel H., Mielsch L. I. and Bornholdt S., 2002, “Scale-Free Topology of E-Mail Networks”, *Physical Review E,* 66, 035103 (R).

Erdős P. and Rényi A., 1959, “On Random Graphs I”, *Publicationes Mathematicae,* 6, 290–297.

Faloutsos M., Faloutsos P. and Faloutsos C., 1999, “On Power-Law Relationships of the Internet Topology”, *Computer Communication Review,* 29, 251–262.

Ferrer-i-Cancho R. and Sole R. V., 2001, “The Small World of Human Language”, *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268(1482), 2261–2265.

Garlaschelli D., Battiston S., Castri M., Servedio V. D. P. and Caldarelli G., 2005, “The Scale-Free Topology of Market Investments”, *Physica A*, 350, 491.

Gorman S., 2005, *Networks, Security and Complexity*, Cheltenham, UK, Edward Elgar.

Gorman S. P. and Kulkarni R., 2004, “Spatial Small World: New Geographic Patterns for an Information Economy”, *Environment and Planning B: Planning and Design*, 31, 2, 273–296.

Gorman S. P., Kulkarni R., Schintler L. and Stough, R., 2000, ”A Network Based Simulation Approach to Cybersecurity Policy”, IMP working paper http://policy.gmu.edu/imp/research.html

Gorman S. P., Patuelli R., Reggiani A., Nijkamp P., Kulkarni R. and Haag G., 2007, “An Application of Complex Network Theory to German Commuting Patterns”, in T. Friesz (ed.), *Network Science, Nonlinear Science and Infrastructure Systems,* Springer-Verlag, New York (forthcoming).

Govindan R. and Tangmunarunkit H., 2000, “Heuristics for Internet Map Discovery”, *Proceedings of IEEE INFOCOM 2000, Tel Aviv, Israel,* (IEEE, Piscataway, N. J.), 3, 1371.

Huberman B. A. and Adamic L. A., 1999, “Growth Dynamics of the World-Wide Web”, *Nature*, 401, 131.

Jeong H., Néda Z. and Barabási A. L., 2003, “Measuring Preferential Attachment in Evolving Networks”, *Europhysics Letters*, 61, 4, 567–572.

Kumar R., Raghavan P., Rajalopagan S. and Tomkins A., 1999, in *Proceedings of the 9th ACM Symposium on Principles of Database Systems, 1.*

Latora V. and Marchiori M., 2001, “Efficient Behaviour of Small-World Networks”, *Physical Review Letters,* 87, 198701, condmat/0101396

Milgram S., 1967, “The Small World Problem”, *Psychology Today,* 2, 60–67.

Newman M. E. J., 2001, “The Structure of Scientific Collaboration Networks”, *Proceedings of the National Academy of Sciences*, USA, 98, 404–409.

Newman M. E. J., 2005, “Power Laws, Pareto Distributions and Zipf’s Law”, *Contemporary Physics,* 46, 323–351.

Nitsch V., 2005, “Zipf Zipped”, *Journal of Urban Economics*, 57, 86–100.

Redner S., 1998, “How Popular is Your Paper? An Empirical Study of the Citation Distribution”, *European Physical Journal B,* 4, 131–134, condmat/9804163

Patuelli R., Reggiani A., Gorman S. P., Nijkamp P., and Bade F.-J., 2007, “Network Analysis of Commuting Flows: A Comparative Static Approach to German Data”, in A. Reggiani, P. Nijkamp (eds.), Special Issue on “Transport Networks and Metropolitan Development: New Analytical Departures”, *Networks and Spatial Economics* (forthcoming).

Russo G., Reggiani A., Nijkamp P., 2007, “Spatial Activity and Labour Market Patterns: A Connectivity Analysis of Commuting Flows in Germany”, *The Annals of Regional Science* (forthcoming).

Schintler L., Gorman S. P., Reggiani A., Patuelli R. and Nijkamp P., 2005a, “Small-World Phenomena in Communication Networks: A Cross-Atlantic Comparison”, in A. Reggiani and L. Schintler (eds.), *Methods and Models in Transport and Telecommunications: Cross Atlantic Perspectives*, Springer Verlag, Berlin, 201–219.

Schintler L., Gorman S., Reggiani A., Patuelli R., Gillespie A., Nijkamp P. and Rutherford J., 2005b, “Complex Network Phenomena in Telecommunication Systems”, *Networks and Spatial Economics*, 5, 351–370.

Vinciguerra S., 2005, “Network models: Methodological Development and Empirical Applications with reference to Telecommunication Sectors”, Master Thesis, Faculty of Statistics, University of Bologna (in Italian).

Watts D. J., 1999, *Small Worlds*, Princeton New Jesey, Princeton University Press.

Watts D. J. and Strogatz S. H., 1998, “Collective Dynamics of ‘Small-World’ Networks”, *Nature,* 393, 440–442.

Zipf G. K., 1949, *Human Behaviour and the Principle of Least Effort*, Addison-Westley Press, Cambridge, MA.

Yook S. H., Jeong H. and Barabási A. L., 2002, “Modeling the Internet’s Large-Scale Topology”, *PNAS*, 99, 21, 13382–13386.

ANNEX A

In this Annex A the graphical representation of the different types of network models (RN, SW and SF networks) will be illustrated.

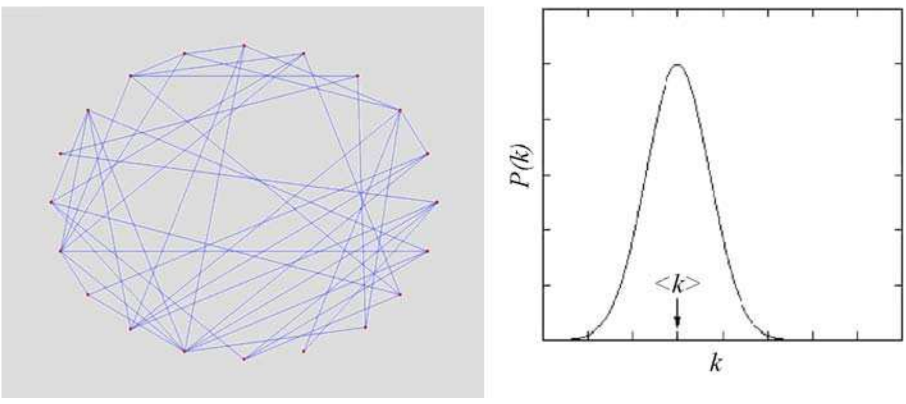


Figure A1. Visualization of a graph representing a RN, simulation made with 20 nodes (left side) (source: elaboration by the authors). On the right side, the Poisson distribution indicates the vertex connectivity degree distribution of a RN (source: Barabási, 1999, p. 105).

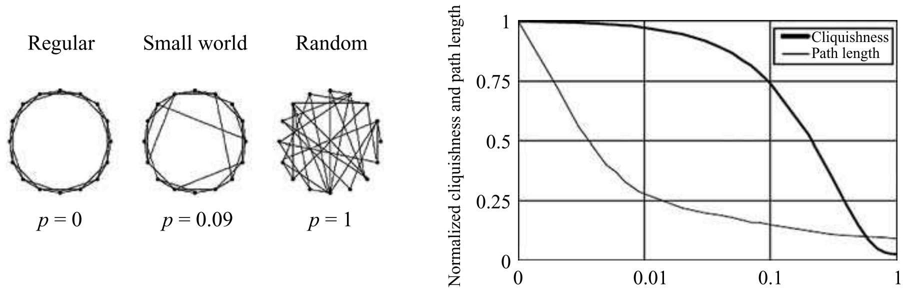


Figure A2. SW networks: transition from a local ordered structure to a RN, dependent of p(left side); the cliquishness (or clustering coefficient) and path length distributions (right side) (source: Cowan, 2004, p. 9–10).

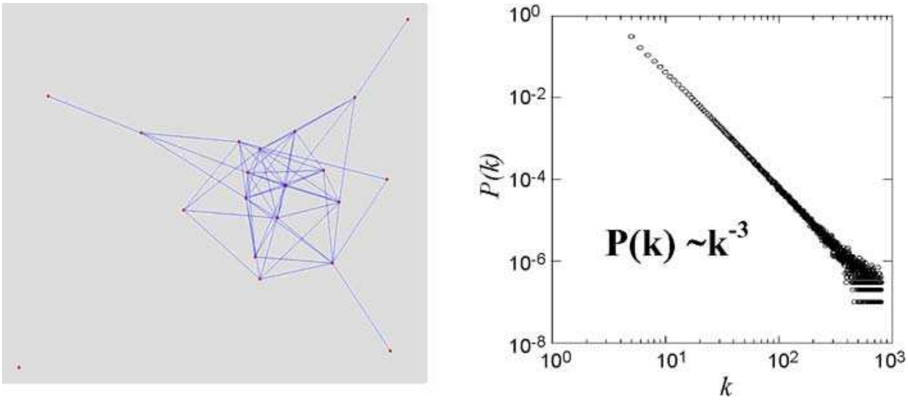


Figure A3. Visualization of a graph representing a SF network, simulation made with 20 nodes (left side) (source: elaboration by the authors). On the right side, the power law distribution indicates the vertex connectivity degree distribution of a SF network (source: Barabási, 1999, p. 511).

ANNEX B

In this Annex B the graphical representation of the log transformation concerning Figures 2–4 (power law vs. exponential law functions) will be illustrated.

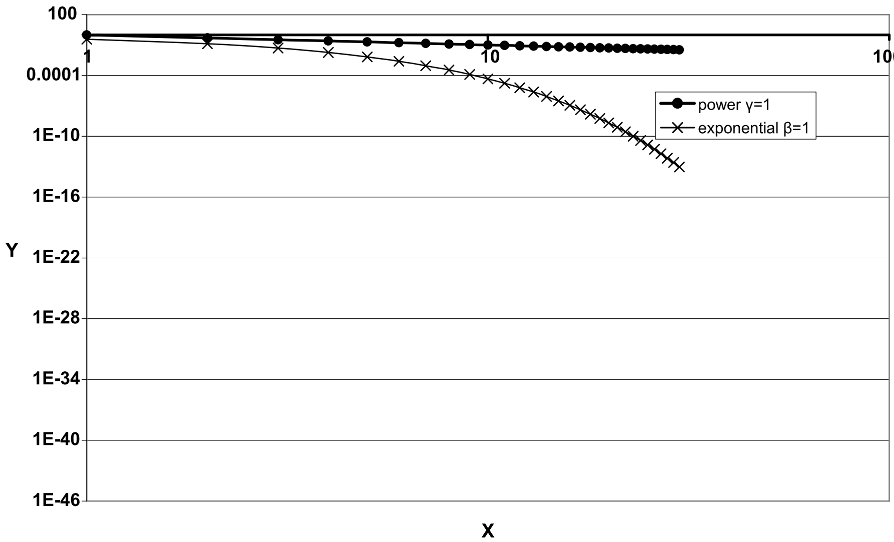


Figure B1. Log-log plot version of Figure 2.

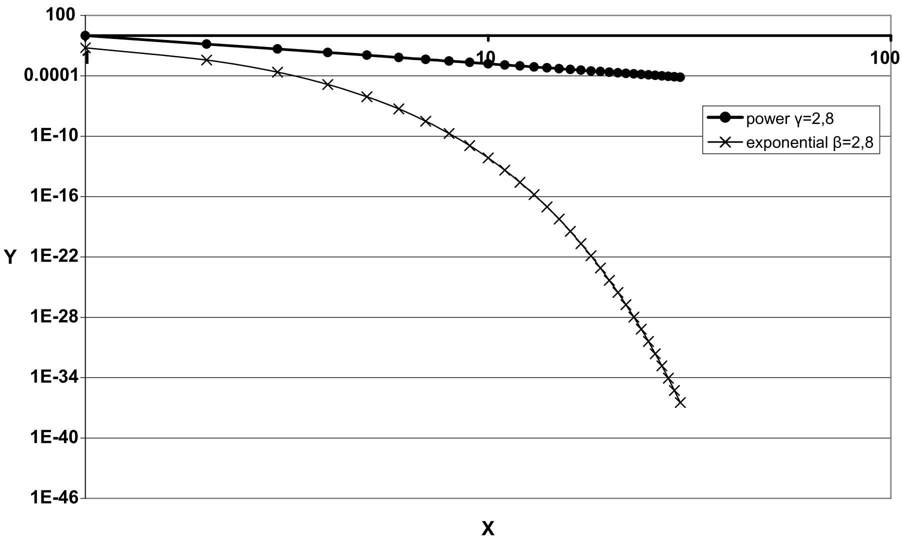


Figure B2. Log-log plot version of Figure 3.

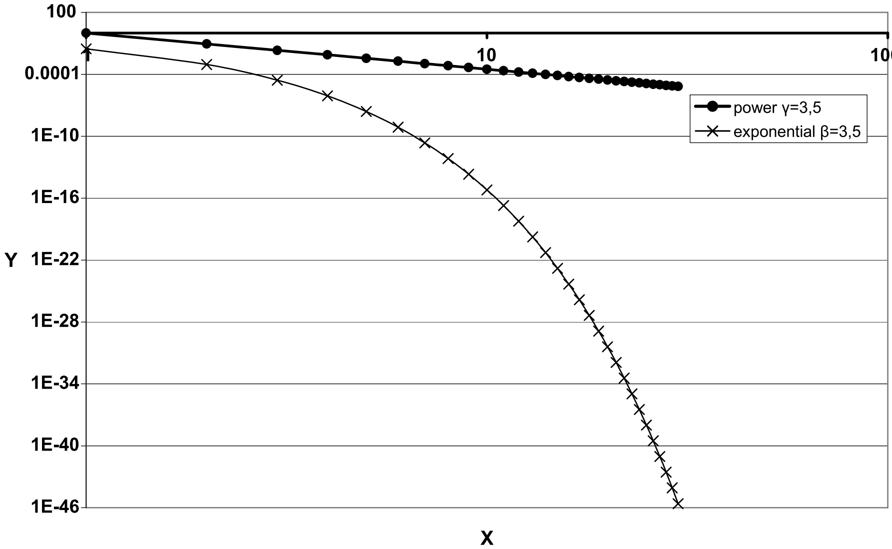


Figure B3. Log-log plot version of Figure 4.

Chapter 8

An Application of Complex Network Theory to German Commuting Patterns

Sean P. Gorman,[1] Roberto Patuelli,[2] Aura Reggiani,[3]
Peter Nijkamp,[2] Rajendra Kulkarni,[1] and Günter Haag[4]

[1] *George Mason University, Fairfax, VA, USA,*
E-mail: sean.gorman@fortiusone.com; rkulkarn@gmu.edu

[2] *VU University Amsterdam, The Netherlands,*
E-mail: rpatuelli@feweb.vu.nl; pnijkamp@feweb.vu.nl

[3] *University of Bologna, Italy, E-mail: aura.reggiani@unibo.it*

[4] *Steinbeis Transfer Centre Applied System Analysis (STASA), Stuttgart,*
Germany, E-mail: haag@stasa.de

Abstract Simulating the structure and evolution of complex networks is an area that has recently received considerable attention. Most of this research has grown out of the physical sciences, but there is growing interest in their application to the social sciences, especially regional science and transportation. This paper presents a network structure simulation experiment utilizing a gravity model to identify interactions embodied in socio-economic processes. In our empirical case, we consider home-to-work commuting patterns among 439 German labour market districts. Specifically, the paper examines first the connectivity distribution of the German commuting network. The paper next develops a spatial interaction model to estimate the structure and flows in the network concerned. The focus of this paper is to examine how well the spatial interaction model replicates the structure of the German commuting network as compared to complex network models. Finally, the structure of the physical German road network is compared to the spatial flows of commuters across it for a tentative supply-demand comparison.

Keywords: complex networks; commuting; infrastructure

1. Introduction

The separation of residential and job decisions have led to complex commuting patterns which have extended in geographical scale over the past decades. As a consequence, home-to-home trips have adopted multi-regional network configurations and have thus led to complex interactive networks. Commuting has become an important field of study in geography, transportation science and regional science (see Rouwendal and Nijkamp, 2004). Commuting has long been studied mostly in terms of forecasting and approximating flows (see, for example, Fotheringham, 1983; White, 1977, 1986). Recent works include the application of models such as the one developed at STASA (Haag et al., 2001). However, less efforts have been done in studying the structure and connectivity properties of commuting networks. A certain amount of literature is available, which studies commuting in a spatial framework. In a recent paper, Cörvers and Hensen (2003) used regional modelling in order to study functional relationships between regions. This approach was carried out in order to improve understanding of commuting behaviour. In particular, the authors' objective was to overcome the limitations of administrative regions in defining new areas that would maximize internal commuting.

A number of works have also been touching on the incorporation of the spatial configuration of commuting destinations, in particular with the works by Fotheringham (see mainly Fotheringham, 1983), who introduced competing destinations models. The competing destinations approach allowed to introduce in the Spatial Interaction Models (SIMs) an element representing the effects of the clustering of destinations, by means of particular accessibility measures. The judgement on this methodology is, however, not uniform. Network approaches to commuting have also been proposed, both at urban level (see, for example, Sheffi, 1985), and at zonal level (Thorsen et al., 1999). A graph theory approach has instead been proposed by Binder et al. (2003).

The above approaches, though, mainly revolve around the analysis of the effects of the road network on commuting. One further level of analysis is therefore necessary, which looks at the behaviour shown in terms of connections between the nodes of the commuting network. Several questions, in fact, need to be answered about the network. Is it highly centralized or decentralized? What are the efficiency and reliability implications of this and other of its connectivity properties? Complex network theory – if applied to commuting networks – can help answering these questions. A wide literature bloomed in recent years, studying the structural and performance implications – on transportation networks – of hypothetical natural disasters or terrorist attacks. The centralization or clustering levels of networks are therefore critical in such discussion.

In this paper, we intend to analyse the network properties of observed home-to-work commuting in Germany. The findings of this analysis are ultimately compared to the ones found by carrying out two simulation models: the first model is an unconstrained SIM, while the second model proposed is modelled according to the scale-free network theories recently made popular by the works of Barabási and Albert (BA) (Barabási, 2001; Barabási and Albert, 1999), showing that networks with preferential attachment-based growth tend to be highly efficient and centralised. In addition to the above analysis, we also propose an analysis of the main German road network, by means of a shortest-path algorithm, and subsequently compare the structural properties found – for the road network – to the ones found for the real data and the simulation models.

The paper is therefore structured as follows: the next section briefly reviews complex network theories and their main implications to our discussion. Section 3 presents a some issues associated with SIMs, and introduces the model carried out for our experiments. Subsequently, Section 4 will first present the results of our empirical application. First, the data used for our analysis are presented, then the findings of the comparison between the properties of the observed commuting network and the simulation models are shown. Finally, a discussion of these findings in respect of the ones obtained by an analysis of the physical German road network is presented. Lastly, conclusions and future research directions are drafted in Section 5.

2. Complex Network Theory: A Brief Review

This section briefly reviews the main issues related to complex network theories, and in particular their implication for transportation networks. Contrary to the attention complex networks have been receiving in recent years, the study of such networks is not particularly new. Before Albert and Barabási's groundbreaking discoveries, original research had in fact been carried out some forty years ago by Erdös and Renyi (ER) (1960), whose major assumption was an underlying random network structure. However, because of lacking computational power and suitable data, for the majority of the 20th century these theories were not adequately challenged and represented the basis for the most common methods of network simulation (Barabási, 2001).

Finally, Albert and Barabási (2002) found, in more recent times, that (large) complex networks were actually behaving according to three main characteristics:

1. Short average path length
2. High level of clustering
3. Power law and exponential degree distributions

In detail, short average-path length indicates that any two nodes on the network can be reached with a limited number of hops. High clustering, instead, occurs because of nodes locating topologically close to each other in cliques that are well connected to each other. This property had been formalised by Watts and Strogatz (1998). Finally, the frequency distributions of node density (or, more generally, number of connections) are called degrees and can follow power-law distributions. This third property implies connections that cut across the graph, directly linking different clusters of vertices. These direct links between clusters bring an increased level of efficiency – in terms of number of hops – to the network. This result shows the limits of the ER models, in which the exponential decay of the degree distribution did not imply a higher number of connections available to the most important nodes.

The novelty in the AB approach was in fact incorporating an additional component: network growth. Consequently, not only the number of nodes in the network can increase, but new nodes are found to have a higher probability of connecting to other nodes that are already well-connected. Formally, the mechanisms that govern network growth towards a power-law degree distribution are (Chen et al., 2001):

a. Incremental growth – As observed above, the number of nodes in the AB models is allowed to grow.
b. Preferential connectivity – Preferential connectivity expresses the frequently encountered phenomenon that new nodes have a higher probability to connect (or reconnect) to an existing node that already has a large number of connections (i.e. high vertex degree).
c. Re-wiring – Re-wiring can be considered as a consequence of the previous principle, as some links can be removed and re-connected in the network, though pointing at new nodes, on the base of preferential connectivity.

Still, after the recent developments described, there is a debate on how complex networks should be classified. Different ideas, for example, are proposed by Albert and Barabási (2002) and Amaral et al. (2000) (see Schintler et al., 2005). Generally, the cause-and-effect relationships underlying large complex networks are still not exactly clear. Furthermore, the measures according to which networks should be measured are also in discussion. For example, Li et al. (2004, p. 11) suggest that – in the case of engineered networks – robustness should be "defined in terms of network performance" and be "consistent with the various economic and technological constraints at work." Remarkably, Li et al. employ a gravity model in generating their network's maximum throughput.

A certain amount of literature is now available on the analysis of transportation networks in terms of complex theory. Because of their short average-path length, airline networks have been considered by Amaral et al. (2000) as a *small-world* network, referring to the model presented by Watts and Strogatz

(1998). On the other hand, the same authors note that structural limitation of airline networks, such as limited space available in the airports, may hinder the emergence of scale-free properties. Other authors found similar results. Latora and Marchiori (2002) analysed the Boston subway network, while Schintler and Kulkarni (2000) observed congested road networks. Both articles found small-world network properties in the analysed networks.

Generally, one might argue that transportation networks are less prone to evolve into a scale-free structure over time given the fact that they tend to be planar. In fact, in planar networks, the maximum number of connections for a single node can be limited by the physical space available to connect it to other nodes, and it is this fact that makes the large number of connections needed for finding a power-law distribution more difficult to obtain. Further, it may be observed that highly centralized transportation networks can be subject to threats to viability, in case of destruction of large hubs (Kwan et al., 2003). Scale-free networks have many implications, but a far-reaching consequence of their unique hub structure is that they are very fault tolerant, while also susceptible to attack (Albert et al., 2000). Specifically, a scale-free network model remains connected when up to the 80% of nodes are randomly removed from the network, but when the most connected nodes are removed, the average path length of the network increases rapidly, doubling its original value when the top 5% of nodes are removed (Albert et al., 2000). In short, targeting the most connected nodes can cause significant damage to a scale-free network, making it highly susceptible to a coordinated and targeted attack. Further, these numbers and findings were highly similar to the ones found when real-world networks were tested, including the Internet at the autonomous system level, and the WWW. When the most connected networks and web pages were attacked, the network rapidly failed. Albert et al.'s work was complimented by the analysis of Callaway et al. (2000), modeling network robustness and fragility as a percolation, and by Cohen et al. (2001), who used related methodologies. Preliminary analyses of these models on spatial network data have shown similar results when cities are the nodes and fiber connections between them are the links. Utilizing a model of node connectivity and path availability, Grubesic et al. (2003) found that the disconnection of major hub cities can cause the disconnection of peripheral cities from the network. Spatial analysis of network failure has also been done for airline networks, finding similar results for the Indian airline network (Cliff et al., 1979).

Starting from these considerations, the next section will present the SIM that was modelled as an approximation of preferential attachment, in order to be compared to a scale-free model inspired by the theories described above.

3. Spatial Interaction Models: An Approximation Tool for Preferential Attachment

3.1 Spatial Interaction Models for Identifying Commuter Flows in the German Labour Market Network

Spatial interaction models are arguably one of the most common methods employed and studied for estimating commuting flows (see, recently, Thorsen and Gitlesen, 1998; Johansson et al., 2003; Jörnsten et al., 2004). Generally, SIMs have long been a popular technique for describing and explaining behavioural, demographic and economic phenomena in space (see Sen and Smith, 1995, for an extensive presentation of the family of methods). The main reason for the widespread utilization of SIMs is their simple mathematical form, in addition to the intuitive assumptions underlying the approach. It should be remembered that the most common specification of SIM had its origins in a resemblance to Isaac Newton's law of universal gravitation. The idea of utilizing models derived from this theory had already been introduced, in the 19th century, in the field of social sciences by Carey (1858) and Ravenstein (1885), and subsequently mathematically formalized by Stewart (1941). Remarkably, SIMs have been shown to have theoretical justification in the entropy theory and in utility maximization/cost minimization (see, for example, Nijkamp, 1975; Nijkamp and Reggiani, 1992). While Isard (1960) first suggested the use of SIMs in regional science, the entropy root of SIMs introduced by Wilson (1967, 1970), and subsequently the micro-economic derivation introduced by McFadden (1974, 1979) contributed to make SIMs more suitable to interpret spatial-economic phenomena.

The common form of an SIM (here presented as double-constrained) is as follows:

$$T_{ij} = A_i B_j O_i D_j f(-\beta c_{ij}) \qquad \text{for} \quad i = 1, \ldots, I; \; j = 1, \ldots, J, \tag{1}$$

where:

$$A_i = 1/\sum_j B_j D_j \exp(-\beta c_{ij}); \tag{2}$$

$$B_j = 1/\sum_i A_i O_i \exp(-\beta c_{ij}); \tag{3}$$

T_{ij} measures the flow of interaction between the origin i and the destination j, depending on the stock variables O_i and D_j, as well as on the deterrence function $f(-\beta c_{ij})$, and on the balancing factors A_i and B_j(see Reggiani, 2004). Generally, Fotheringham and O'Kelly (1989, p. 10) formulate a SIM in the general framework of the Alonso model (1978) as follows:

$$T_{ij} = f(\mu_1 v_1^1, \mu_2 v_1^2, \ldots, \mu_p v_1^p; \alpha_1 w_j^1, \alpha_2 w_j^2, \ldots, \alpha_q w_j^q; \beta c_{ij}), \tag{4}$$

where v_i and w_j are measures of the "propulsiveness" and the "attractiveness" of i and j, respectively. Parameters μ, α and β link the above variables to the flows T_{ij}.

The deterrence function in (1) is depending on the deterrence factor β and the interaction costs c_{ij}. c_{ij} might also be considered as generalized costs. In our experiment, distances were used as a proxy of the interaction costs, since the analysis was carried out at the German district level (*kreise*). The functional form of the deterrence function is also a relevant issue. While in its first formulations the distance deterrence function was shaped as a power law function – as used in the Newtonian formula – Kulldorf (1955) showed that an exponential deterrence function seemed to better fit migration phenomena. Subsequently, the exponential deterrence form emerged mathematically from the entropy maximization approach developed by Wilson (1967). The power form outlines a larger amount of flows – with respect to the exponential form – in the presence of long distances or travel times. In our analysis, the power law specification was used, as it showed to fit the data better. In addition to the shape of the deterrence function, the value of the β deterrence factor was researched. In the experiments conducted here, a value of 1.5 was chosen for the β deterrence factor, on the basis of a calibration procedure carried out on the available data expressed in the form of an unconstrained SIM. In particular, the unconstrained SIM used in our experiments is specified as follows:

$$T_{ij} = K E_i E_j d_{ij}^{-\beta} \tag{5}$$

In Equation (5), the flows T_{ij} are the employees commuting from the origin district i to the destination district j. They are a function of the number of persons E_i and E_j employed in the two districts, as well as of the distance d_{ij} between the two, in addition to a scaling factor K. The model that we propose is of course overly simple. However, what is relevant for our experiments is not the correct estimation of the German commuting flows, but instead the connectivity and structure of the commuting network (see Section 3.2).

When employing an SIM for estimating inter-urban commuting flows, additional issues should be cited. One of them is the treatment of internal commuting. In particular, the distance between the working and living areas is, by definition, null (although travel time or costs would not necessarily be). This issue is at times solved by assigning an arbitrary value to the distance for internal commuting. Alternatively, the flows assigned to internal commuting can be omitted in the analyses. A number of additional ways to treat internal commuting are available in the literature. The method suggested by Thorsen and Gitlesen (1998) starts from the consideration that intra-commuting might imply different transportation means, such as biking or walking. Thorsen and Gitlesen suggest an additional component to be added to the deterrence function exponent. This component would represent – depending on the case – either a start-up (generalized) cost for commuting between different zones, or a premium, interpreting the benefit of intra-commuting. An example model with these characteristics, reminiscent of the Champernowne deterrence function (see, for example, Sen and Smith, 1995), is presented by Thorsen and Gitlesen (1998, p. 279) for a double-constrained specification. Alternatively, the authors suggest that labour market characteristics might be used to influence the elements on the diagonal of the O/D matrix.

In our case, the elements of the diagonal are omitted from the analysis. This choice was made mainly due to the network approach to commuting identified in the paper. As we analyse the connectivity and structural properties of the German commuting network, the measure of the number of commuters within a certain district would not add additional information about the network, apart from the "socio/economic weight" of a certain node. On the other hand, the number of fulltime employees in each district already grasps this aspect.

3.2 Interpretation of Spatial Interaction Behaviour as Preferential Attachment

The usual practice in the use of SIMs, when dealing with commuting flows, is to employ the models in forecasting future flows, given certain conditions. In our experiments, we propose the utilization of the simple SIM shown in Equation (5) as a tool for approximating the connectivity and structural properties of a commuting network. In particular, we want to verify if an SIM can allow for preferential attachment behaviour. As seen in Section 2, in the models introduced by Barabási and Albert nodes have a higher probability of connecting to other nodes that are already well-connected. The hypothesis that we will test in the next section is that commuting networks follow a similar preferential attachment-based behaviour in terms of connectivity and structure. They would not be the first transportation network to be referred to in these terms. In fact, hub-n-spoke networks operated by airlines are, as seen in Section 2, a well-known example of preferential attachment behaviour (see, for example, Bowen, 2002, and most importantly Wojahn, 2001).

An additional reason for the consideration of commuting networks in such a framework can be found if we think of preferential attachment as a maximization of utility levels. The idea is that utility is maximized by connecting to the most connected nodes of the network, as they give access to other points in the network by a minimal number of hops (therefore minimizing generalized costs). If so, this hypothesis would be consistent with the theoretical basis of utility maximization that justifies the use of SIMs. In particular, the hub-n-spoke network might – conceptually – be interpreted as a network tree consistent with a nested logit/hierarchical SIM structure (for the compatibility between nested logit and double constrained SIM, see Nijkamp and Reggiani, 1992).

The next section will first describe the data available for the experiment (see Section 4.1). We will then test the hypothesis of a SIM as an approximation of preferential attachment, by comparing the network properties observed for our naïve SIM, for a scale-free network, and for the real commuting network (Section 4.2). Subsequently, in Section 4.3 we will present the results of a structural analysis of the German road network (the physical network on which commuting is actually performed), and compare the properties found with the results previously obtained.

4. The Empirical Analysis

4.1 The Data Available

As shown in Section 3.1, the SIM estimated for our experiment formally employs two types of data:

a. The number of employees working in each German district. These are fulltime employees, for which the data were collected – as part of a yearly survey – in the year 2002.
b. The distance between each couple of origins and destinations. This is expressed in Kilometres, and acts as a proxy for more effective measurements, such as cost or travel time.

Moreover, additional data are available for our experiment. In order to calibrate our SIM shown in model (5) (see Section 3.1), information on the German commuting flows has been used. The data consists, for each origin-destination couple (i, j), of the number of employees living in district i, and working in district j. These are therefore home-to-work data, which are available for the year 2002. The distribution of the number of commuters for each origin-destination couple, ranked in descendent order, is shown in Figure 1.

Figure 1, which is adapted to a log-log scale, shows the decrease in the number of commuters for the (descendent) rank of origin-destination couples. The curve seems to better fit a power-law distribution rather than an exponential one. However, the highest values in the data – which are the first values in the descending order rank – do not reach the levels expected for a power law fitting. This could reasonably be due to physical constraints given by city size

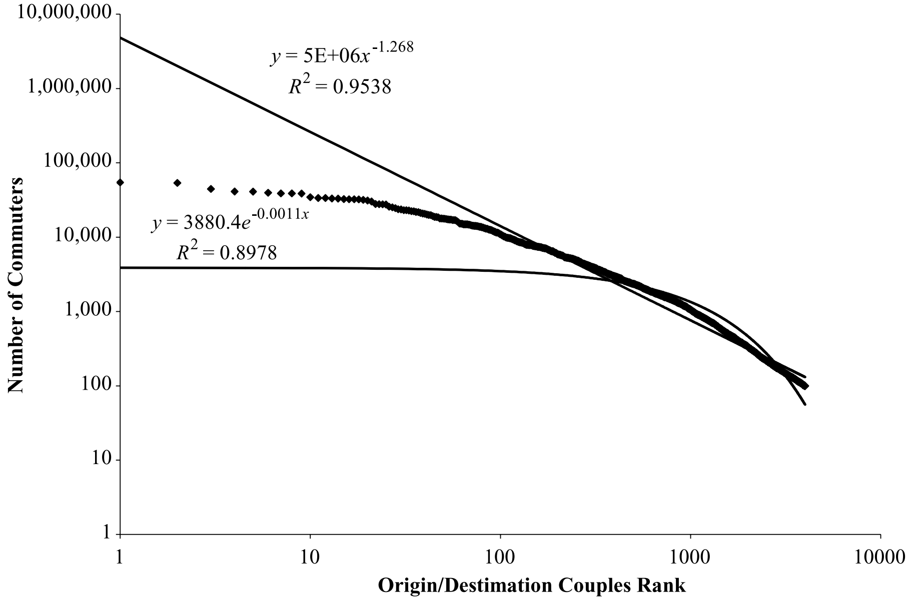


Figure 1. The distribution of commuters flows over the couples of origins and destinations.

and congestion issues. Nevertheless, the distribution shown in Figure 1 does not tell about the intrinsic properties of the network. The next section will therefore analyse the connectivity properties of the observed commuting network and of the ones developed by the models presented in our exercise.

4.2 Analysis of Network Structure and Connectivity Properties in Germany

The goal for this experiment is to discover how well a spatial interaction model can be used to accurately model preferential attachment and the resulting structure of the German commuting network. The use of the BA model as a structural model for network connectivity has received considerable attention, critique and extension. This section of the paper strives to refine the BA model for application to spatial economic networks, specifically the patterns of connectivity found in German commuting to work. The focus of the analysis is how well the spatial interaction model builds the topology on the real world network and not just its connectivity distribution. Li et al. (2004) have found that the power-law distributions created by many network generation models can result in a wide range of actual topologies.

The first step of the analysis was to specify an accurate SIM for the German commuting network. Many economic and social forces shape the

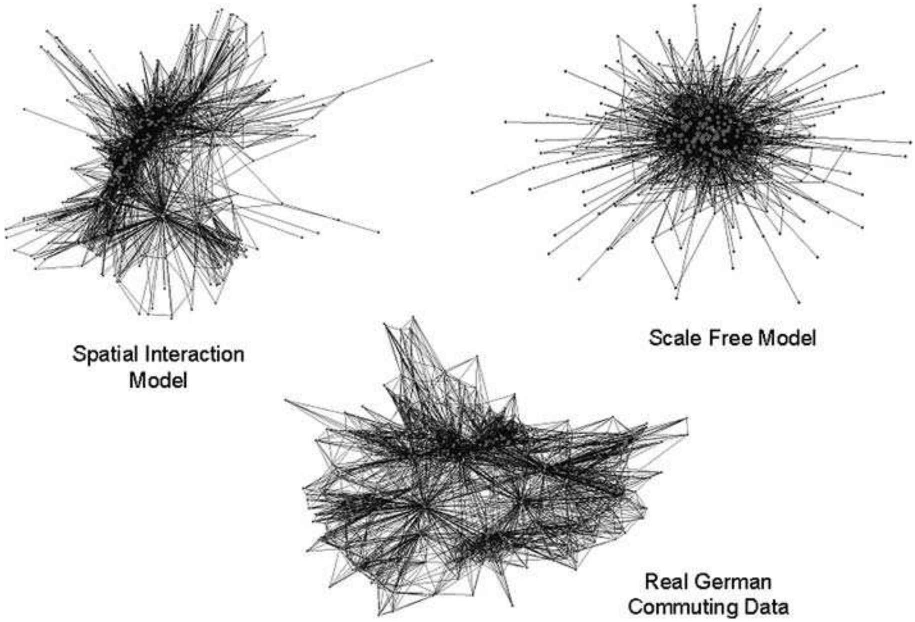


Figure 2. Network Visualization of the Network BA Model, Spatial Interaction Network Model, and German Commuting Network.

contours of commuting patterns. A simple supply-and-demand model based on employment and distance was used in this initial case. The SIM was formalized in Section 3.1, Equation (5). Once the spatial interaction for all possible pairs of German labor markets is calculated, a threshold for connectivity is determined. In this case, a threshold of 100 commuters was determined as the cut off for what would be considered an adequate economic flow for there to be a connection in the network between two (labor) districts. The resulting edge list of pairwise connections between districts was then used for comparison to the BA model and real-world data. The BA model created for comparison was based on a 439-node network with a 0.3 connectivity probability, an alpha parameter of 0.3, and an initial three districts to connect to. Once both networks are modeled, they are compared to each other and the real commuting network structure. The resulting network topologies can be seen in Figure 2.

From a visual inspection of the three network visualizations, the spatial interaction network model comes closest to replicating the German commuting network, although it lacks the same level of interconnectivity seen in the real data. The BA model illustrates even less interconnectivity than the spatial interaction model, with most connections going directly to the hubs, with little of the connectivity between spokes seen in the real commuting data.

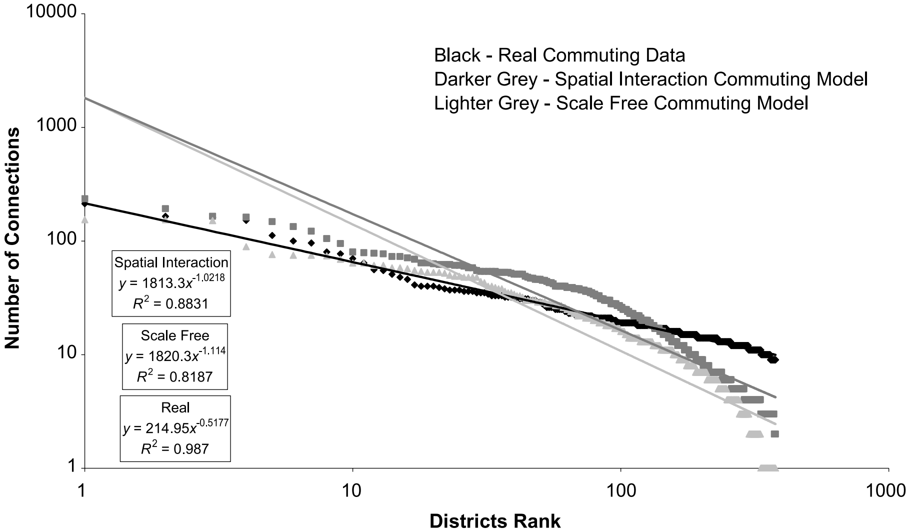


Figure 3. The resulting connectivity distributions of the BA network model, spatial interaction network model, and the German commuting network.

In order to provide the next level of analysis the connectivity distribution of each network is calculated as a log-log plot. The results of all three network connectivity distributions are plotted in Figure 3. The number of connections for each district is ranked in descendent order.

Interestingly, neither of the models recreated the power-law connectivity distributions seen in the real data. They have instead exponential distributions of connectivity. The two models also have much steeper slopes, each with an exponent over 1 while the real data is just over 0.5. Of the two, the SIM is modestly closer in slope and distribution to the real data, but not enough to be of consequence. Overall, the three distributions are highly similar despite the slight numeric differences, while the topology visualizations seen in Figure 2 are vastly different. This is most prominent in the SIM and the BA model, which have very similar connectivity distribution, but entirely different topology visualizations. To examine these topologic differences, a variety of structural indicators are calculated.

For each network, a series of indicators is calculated, which provide a comparison of the structure of each network generated (see Table 1).

The first indicator presented in Table 1 is the clustering coefficient, which provides the first insight into what delineates the structural differences in the three networks. The lack of clustering seen visually in the BA model is found in the statistics, with a far lower coefficient than that of the SIM or of the real world data. The SIM overestimates the level of clustering seen in the real data, but is considerably closer than the BA model. The diameter

Table 1. Network properties of the BA network model, spatial interaction network model, and the German commuting network.

	Real Data	BA Model	SIM
Clustering Coeff.	0.659	0.29	0.803
Diameter	4	6	5
Average Degree	18.21	12.469	18.462
Std. Dev. Degree	19.268	19.174	26.805
Max Degree	213	154	235
Min Degree	5	0	1
Centralization	44.68%	32.46%	50.36%
Betweeness Mean	313.125	251.866	297.083
Betweeness Std. Dev.	1929.005	935.554	1997.69

parameter calculates the longest shortest-path in the network, and provides a measure of the efficiency of the structure. All three networks have relatively small diameters, and the real world commuting network is the most efficient. Again, the diameter of the SIM is closer than the one of the BA model, off by only 1. The subsequent set of statistics deals with the number of connections nodes in the network have. These are the statistics that are more of local connectivity indicators, and not global structural indicators of the networks. From a connectivity standpoint, the SIM does a nice job of accurately capturing the average, maximum, and minimum of the degree connectivity seen in the real commuting data. It falls short in capturing the standard deviation of average connectivity, where the BA model is closer, but the average connectivity is still six degrees off of the real data. The centralization parameter measures the amount of core connectivity in the network. The SIM overestimates the amount of centralization, while the BA model underestimates it. However, the SIM is closer to the real data values. Lastly, a measure of betweenness is offered for each network. Betweenness is a measure of routing frequency, where all shortest paths across a network, and then the number of times each node is used in all paths, are calculated. This provides a convenient measure of the global structure of the network, since the indicator samples paths across all segments of the network. The average and standard deviation of betweenness for the SIM and the real data are very close. This confirms what was seen previously in the network visualization, which illustrated a similar topological structure between the two networks. The significant differences between the real network and the BA model, according to betweenness measures, confirm the distinct differences between the two also seen in the visualization.

4.3 A Structural Analysis of the Physical Commuting Network by Means of Shortest Path Algorithm

In addition to studying the flows of commuters between cities, it is also possible to study the infrastructure they utilize. The economic flows of commuters and the physical links of infrastructure are intrinsically connected, but belong to two very different network structures. Commuting flows belong to logical networks, which are non-planar in nature, since the fact that two links intersect does not mean a node actually exists at their intersection. A flow in the commuter network could therefore be between Frankfurt and Munich with only two nodes and one link, even though the physical path goes through Stuttgart. The physical network, on the other hand, is planar; the intersection of two links creates a navigable intersection. In order to travel from Munich to Frankfurt, several intermediate nodes have to be traversed. Commuting data represent the flows across the physical network, but the two networks are quite different in nature and structure.

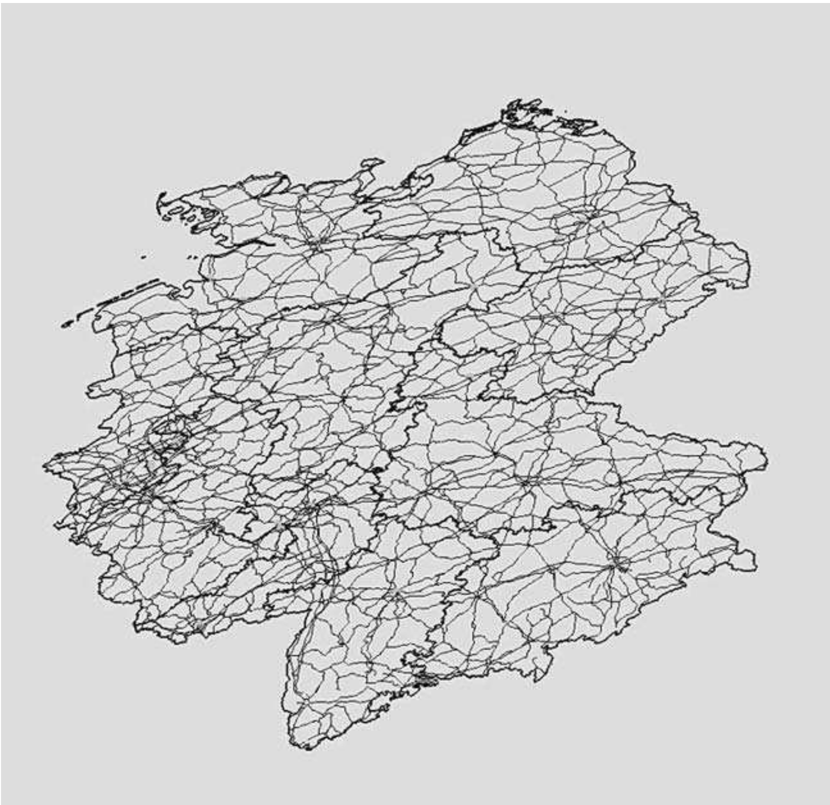


Figure 4. The German road network.

To begin to address the relationship and differences of these networks, the physical road network of Germany is analysed. Unlike the commuting flow network, it is straightforward to visualize what the road network looks like with a simple map. Figure 4 provides a map of the German road network.

In Figure 4, the white lines are major roads in Germany, while the dots represent major cities connected by the roads. While the map does illustrate the layout of the road network, it does not give much insight into its structural properties. In order to gain some perspective on the structure of the road network, a routing frequency analysis was performed. The road network was first partitioned into nodes and links, then shortest-paths were calculated to and from all nodes in the networks. Links were then assigned a frequency count, based on the number of times the link was utilized in all possible link combinations. This provided a structural analysis of which links, in the German road network, are most critical and heavily utilized in all possible travel combinations. To visualize these results Figure 5 was built.

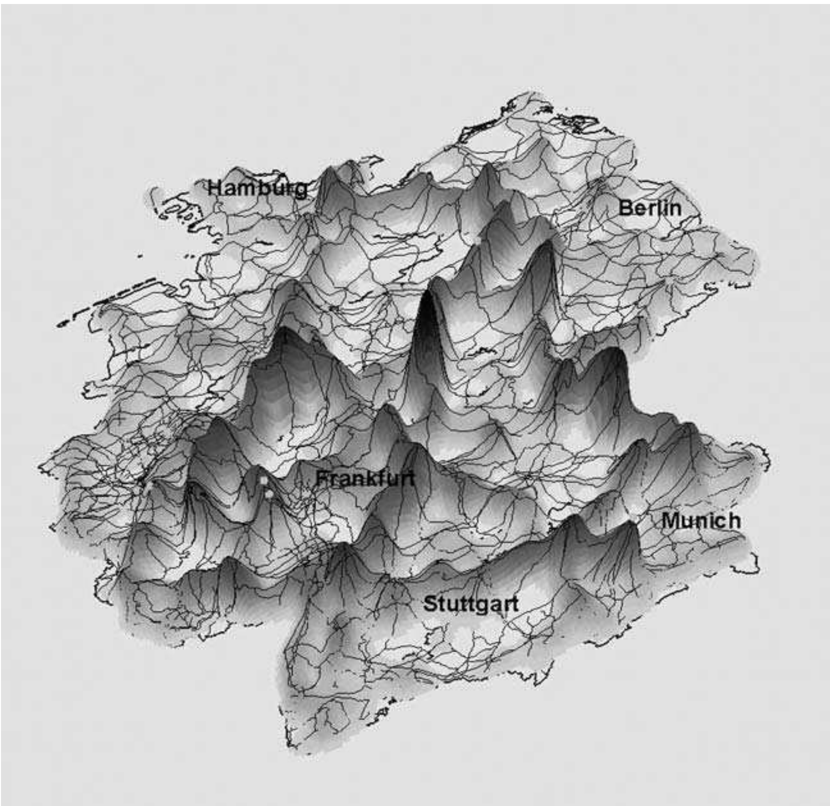


Figure 5. German road network routing frequency analysis.

In Figure 5, the height and colour (lighter to darker) of the peaks is determined by the number of routes that use a particular link in the road network. The higher the frequency of routes, the higher the peak. The routes that connect through the middle of the country are particularly well utilized, especially the routes connecting Berlin to Frankfurt and Stuttgart. The routes connecting Berlin to Munich and Hamburg are also prevalent. In general, routes in Western Germany have a higher frequency than Eastern Germany routes. It should be noted that this analysis is simply based on shortest-path frequency and does not account at all for socio-economic dimensions like population or employment as with the previous gravity model.

4.4 Concluding Remarks

The present analysis does provide preliminary insights into the German infrastructure network, which underpins the economic commuter network flowing on top of it. The intuitive next direction is to combine the two networks and examine their interdependencies. It would be useful to investigate the relationship between the economic flows of commuters and the physical network structure of the roads they use. Are the highest flows of commuters also utilizing the highest-frequented structural links? How well does the physical structure of the network match the economic flows across it? If there is failure in the physical network, how will it impact the economic flow of commuters or, more importantly, logistics and supply chain networks? Unfortunately, it was not possible, at this stage, to obtain a geo-referenced map of the German districts to map the commuter flows on top of the physical routing frequency. Such a possibility would at least provide a first-cut comparison of commuting links in a labour district to the number of routing paths. These are all possible future directions for the research.

5. Conclusions

The aim of this paper was to provide a comparative analysis among different approaches on the theme of commuting from a network perspective. In particular, we were interested in studying the structure and properties of the German commuting network. In order to better understand this real-world commuting network, we needed to compare it to different network models that could approximate it. Two models (an SIM and a scale-free model) were developed, based on widely different assumptions. However, both models aimed at simulating a “preferential connectivity” behaviour, according to which well-connected nodes in the network have a higher probability of attracting connections from other nodes over the network.

A comparison of the real (network) commuting data with the ones generated by our two models was carried out in two ways: a) a visual comparison of the network structures; b) a comparison of the network properties calculated for the three networks. The visual comparison showed that the SIM seemed to better approximate the decentralized configuration of the commuting network. The scale-free network showed instead a highly centralized structure. These observations were then reinforced by the comparison of the network properties parameters. Although showing similar values for some network parameters, the SIM, rather than the scale-free model, provided values that were more consistent with the ones calculated for the real data.

An additional analysis was subsequently carried out, examining the German physical road network. This analysis on the road infrastructure visually showed which points, according to a shortest-path routing algorithm, are (theoretical) critical points in the German road network, as they are central to the routes calculated over the network.

Summarizing, our experiments made a first attempt at interpreting commuting networks from a complex network perspective. More detailed experiments might be carried out, by developing more refined SIMs (like double-constrained SIMs) and scale-free models, which should include parameters that better suit the type of network that is being approximated. A further experiment might be to weigh the number of connections – in a scale-free approach – with the volume of commuters. Also, a set of new experiments could be carried out, by integrating commuting flows (and maybe other economic factors) with the physical road network, as suggested in Section 4.4. For example, it could be interesting to observe how changes in the physical road network would influence the results of a SIM, where the distance deterrence factor is not expected to vary (Jörnsten et al., 2004). Also, the use of double-constrained SIMs is necessary, in order to account for spatial characteristics. General equilibrium models, employing variables such as wages or migration might be used for comparison sake.

References

Albert, R., Barabási, A.-L. (2002). Statistical Mechanics of Complex Networks. *Review of Modern Physics,* 74, 47.

Albert, R., Jeong, H., Barabási, A.-L. (2000). Error and Attack Tolerance in Complex Networks. *Nature,* 406, 378–382.

Alonso, W. (1978). A Theory of Movements. In N. M. Hansen (Ed.), *Human Settlement Systems*. Cambridge, Mass.: Ballinger.

Amaral, L. A. N., Scala, A., Barthélémy, M., Stanley, H. E. (2000). Classes of Small-World Networks. *Proc. Nat. Acad. Sci. USA,* 97, 11149–11152.

Barabási, A.-L. (2001). The Physics of the Web. *Physics World,* July 2001.

Barabási, A.-L., Albert, R. (1999). Emergence of Scaling in Random Networks. *Science,* 286, 509–512.

Binder, J., Haag, G., Rabino, G. (2003). Analysis and Modelling of Commuter Flows: Application to the Regions of Stuttgart and Turin. *Jahrbuch der Regionalwissenschaften,* 23, 117–139.

Bowen, J. (2002). Network Change, Deregulation, and Access in the Global Airline Industry. *Economic Geography,* 78(4), 425–439.

Callaway, D. S., Newman, M. E. J., Strogatz, S. H., Watts, D. J. (2000). Network Robustness and Fragility: Percolation on Random Graphs. *Physical Review Letters,* 85(25), 5468–5471.

Carey, H. C. (1858). *Principles of Social Science*. Philadelphia: J. Lippincott.

Chen, Q., Hyunseok, C., Govindan, R., Sugih, J., Schenker, S., Willinger, W. (2001). *The Origin of Power Laws in Internet Topologies Revisited.* Paper presented at the IEEE Infocom, 2002.

Cliff, A., Haggett, P., Ord, H. (1979). Graph theory and geography. In R. Wilson, L. Beineke (Eds.), *Applications of Graph Theory* (Vol. 293–326). London: Academic Press.

Cohen, R., Erez, K., Ben-Avraham, D., Havlin, S. (2001). Breakdown of the Internet Under Intentional Attack. *Physical Review Letters,* 86(16), 3682–3685.

Cörvers, F., Hensen, M. (2003). *The Regionalization of Labour Markets by Modelling Commuting Behaviour.* Paper presented at the 43rd ERSA, Jyvaeskylae.

Erdos, P., Renyi, A. (1960). *On the Evolution of Random Graphs* (Vol. 5): Publication of the Mathematical Institute of the Hungarian Academy of Science.

Fotheringham, A. S. (1983). A New Set of Spatial-Interaction Models: The Theory of Competing Destinations. *Environment and Planning A,* 15, 15–36.

Fotheringham, A. S., O'Kelly, M. E. (1989). *Spatial Interaction Models: Formulations and Applications*. Dordrecht: Kluwer Academic Publishers.

Grubesic, T. H., O'Kelly, M. E., Murray, A. T. (2003). A Geographic Perspective on Telecommunication Network Survivability. *Telematics and Informatics,* 20(1), 51–69.

Haag, G., Binder, J., Koller, M. (2001). *Modellgestützte Analyse zur Regionalen Entwicklung von Beschäftigungsvolumen, Lohnsummen und Beitragseinnahmen*. Nuernberg and Stuttgart: Institute of Labour and Employment Research (IAB) and STASA.

Isard, W. (1960). *Methods of Regional Analysis*. Cambridge: MIT Press.

Johansson, B., Klaesson, J., Olsson, M. (2003). Commuters' Non-linear Response to Time Distances. *Journal of Geographical Systems,* 5(3), 315–329.

Jörnsten, K., Thorsen, I., Ubøe, J. (2004). Replication/Prediction Problems in the Journey to Work. *Environment and Planning A,* 36(2), 347–364.

Kulldorf, G. (1955). Migration Probabilities, *Lund Studies in Geography, Series B: No. 14*. Lund, Sweden: Department of Geography, Lund University.

Kwan, M.-P., Murray, A. T., O'Kelly, M. E., Tiefelsdorf, M. (2003). Recent Advances in Accessibility Research: Representation, Methodology and Applications. *Journal of Geographical Systems,* 5, 129–138.

Latora, V., Marchiori, M. (2002). Is the Boston subway a small-world network? *Physica A,* 314, 109–113.

Li, L., Alderson, D., Willinger, W., Doyle, J. (2004). *A First-Principles Approach to Understanding the Internet's Router-Level Topology*, *ACM SIGCOMM Computer Communication Review*, 34(4), 3–14.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105–142). New York: Academic Press.

McFadden, D. (1979). Qualitative Methods for Analysing Travel Behaviour of Individuals: Some Recent Developments. In D. A. Hensher, P. R. Stopher (Eds.), *Behaviour Travel Modelling* (pp. 279–318). London: Croom Helm.

Nijkamp, P. (1975). Reflections on Gravity and Entropy Models. *Regional Science and Urban Economics,* 5, 203–225.

Nijkamp, P., Reggiani, A. (1992). *Interaction, Evolution and Chaos in Space*. Berlin Heidelberg: Springer-Verlag.

Ravenstein, E. G. (1885). The Laws of Migration. *Journal of the Royal Statistical Society,* 48(2), 167–235.

Reggiani, A. (2004). Evolutionary Approaches to Transport and Spatial Systems. In P. R. Stopher, K. Button, K. Haynes, D. A. Hensher (Eds.), *Handbook of Transport Geography and Spatial Systems* (pp. 237–253). Amsterdam: Elsevier.

Rouwendal, J., Nijkamp, P. (2004). Living in Two Worlds: A Review of Home-to-Work Decisions. *Growth and Change,* 35, 287–303.

Schintler, L. A., Gorman, S. P., Reggiani, A., Patuelli, R., Nijkamp, P. (2005). Scale-Free Phenomena in Communications Networks: A Cross-Atlantic Comparison. In A. Reggiani, L. A. Schintler (Eds.), *Methods and Models in Transport and Communications. Cross Atlantic Perspectives* (pp. 201–220). Berlin: Springer-Verlag.

Schintler, L. A., Kulkarni, R. G. (2000). The Emergence of Small World Phenomenon in Urban Transportation Networks: An Exploratory Analysis. In A. Reggiani (Ed.), *Spatial Economic Science: New Frontiers in Theory and Methodology* (pp. 419–434). Berlin: Springer.

Sen, A., Smith, T. E. (1995). *Gravity Models of Spatial Interaction Behavior*. Heidelberg and New York: Springer.

Sheffi, Y. (1985). *Urban Transportation Networks*. Englewood Cliffs, NJ: Prentice-Hall.

Stewart, J. Q. (1941). In Inverse Distance Variation for Certain Social Influences. *Science,* 93, 89–90.

Thorsen, I., Gitlesen, J. P. (1998). Empirical Evaluation of Alternative Model Specifications to Predict Commuting Flows. *Journal of Regional Science,* 38(2), 273–292.

Thorsen, I., Uboe, J., Naelig;vdal, G. (1999). A Network Approach to Commuting. *J Regional Sci,* 39(1), 73–101.

Watts, D. J., Strogatz, S. H. (1998). Collective Dynamics of Small-World Networks. *Nature,* 363, 202–204.

White, M. J. (1977). A Model of Residential Location Choice and Commuting by Men and Women Workers. *Journal of Regional Science,* 17(1), 41–52.

White, M. J. (1986). Sex Differences in Urban Commuting Patterns. *The American Economic Review,* 76, 368–372.

Wilson, A. G. (1967). A Statistical Theory of Spatial Distribution Models. *Transportation Research,* 1, 253–269.

Wilson, A. G. (1970). *Entropy in Urban and Regional Modelling*. London: Pion.

Wojahn, O. W. (2001). Airline Network Structure and the Gravity Model. *Transportation Research: Part E: Logistics and Transportation Review,* 37(4), 267–279.

Chapter 9

Assessing Critical Components in Transportation Systems: Economic Models and Complex Network Science Approaches

Satish V. Ukkusuri and José Holguín-Veras

Rensselaer Polytechnic Institute, E-mail: ukkuss@rpi.edu, jhv@rpi.edu

Abstract This paper reviews and discusses modeling approaches to identify critical components of transportation systems. The review includes approaches based on economic theory, complex network science techniques and network optimization models. The economic model derives the criticality of the components of the transportation system using concepts from welfare economics. To derive approximate insights into assessing critical components, a model based on complex network science is developed. This model uses the shortest distance as a measure for the efficiency of the network with and without the components. Finally, the paper discusses network interdiction models, which are useful in identifying critical links under strategic behavior of agents. The modeling methodologies presented here are a promising step in assessing critical components and in the optimal use of scarce funding to improve transportation security.

Keywords: transportation; complex networks; transportation economics

1. Introduction

The tragic events of September 11th, 2001 catapulted transportation security to the forefront of issues. In this context, policy makers and researchers have to develop strategies and policies to protect the nation's transportation infrastructure against external threats such as terrorism. At the same time, the transportation system has to deal with unparallel growth in demand across all modes, both passengers and freight. This poses a significant challenge because security measures tend to restrict mobility and trade. In this context, assessing the importance of a given component of the network provides important guidance in identifying the potential benefits of alternative strategies to increase security. Among other things, it also helps maximize the effectiveness of scarce resources. The critical assessment of different components of the multi-modal infrastructure network (e.g., bridges, ports, highways, airports) has not received significant study in the past. Nevertheless, a number of policy measures have been issued by the Federal government (e.g., the Customs Trade Partnership Against Terrorism, Container Security Initiative, 24 hours manifest rule, the 2002 Maritime Transportation Act) in an attempt to provide an across the board increase in network security. However, these policies do not provide any guidance to local and state agencies on how to prioritize investments aimed at increasing transportation network security.

Executive Order 13010, issued by President Clinton on July 15, 1996, (Clinton, 1996) established the President's Commission on Critical Infrastructure Protection (PCCIP) in order to develop a national strategy for protecting these infrastructures from various threats and to assure their continued operation. Eight different types of infrastructure, including transportation, were identified as critical, in that their incapacitation or destruction would have a debilitating effect on the nation's defense or economic security. Presidential Decision Directive 63 (PDD63) issued on May 28, 1998, (Clinton, 1998) builds on the PCCIP's October 1997 recommendations. The report called for a national effort to assure the security of the United States increasingly vulnerable and interconnected infrastructures. PDD63 set up a new institutional structure to deal with this important challenge. The interdependences among critical transportation infrastructure systems calls for a concerted modeling effort to: (i) better characterize the inherent risks and (ii) to prioritize the existing infrastructure systems to manage them safely. This paper specifically deals with the latter, i.e., assessing the importance of components in transportation systems.

The criticality of a component of the transportation network can be studied using different modeling techniques. The potential approaches range from techniques based on economic evaluation of projects; to techniques based on complex network theory, which provides useful approximations to the evaluation.

The paper conducts a critical examination of alternative approaches to model transportation security related problems. The paper starts with a conceptual discussion of the economic paradigm traditionally used to assess the economic value of component of the transportation infrastructure. Using this paradigm as the starting point, the paper then discusses network models and models based on complex network theory that may provide useful approximations to the economic model, and complementary views to the analysis of transportation security problems. The final section of the paper concludes with thoughts on research opportunities in this subject area.

2. Assessing Criticality in Transportation Networks

Assessing the importance of a component of the transportation network requires understanding of the economic value to its users, and society at large. The essence of this process is best understood as a special kind of economic analysis in which the objective is to assess the economic value of *losing* the facility by a terrorist attack, for instance; as opposed to the more traditional objective of assessing the economic value of *adding* capacity to the transportation network. This section focuses on providing a conceptual discussion of the multi-layered nature of the user types, their relationships with the physical network, and alternative methods to assess criticality.

Computing the economic value of a critical component of the infrastructure system invariably leads to welfare economics, and to the different metrics to assess consumer surplus. For completeness sake, a brief review of key concepts is provided next. In accordance to welfare economics, the economic value of a good is defined by the economic benefits it brings to both consumers and producers. In its simplest (Marshallian) form (after Marshall, 1924; though it was originally inspired by Dupuit in the XIV Century, reprinted in Dupuit, 1952), estimates of the economic benefits to consumers (consumer surplus) and to producers (producer surplus) could be obtained by integration of the areas between the demand function and the equilibrium point, and the area between the equilibrium point and the supply curve. In general:

Consumer surplus:

$$CS = \int_{P_e}^{\infty} Q dp \tag{1}$$

Producer surplus:

$$PS = PQ - \int_{0}^{Q_e} m dQ = PQ - C(Q) \tag{2}$$

Figure 1 (a) represents the case of a good with a demand function D and supply function S_0. As shown, the consumer surplus represents the net difference between the amount the users are willing to pay and what they actually pay, i.e., the area above the equilibrium price and below the demand function, is equal to CS_0; while the producer surplus, defined as the difference between the price at which producers are willing to produce and the current market price, is equal to PS_0. Figure 1 (b) of the figure shows the impact of a capacity reduction that moves the supply function to the left of S_0 to S_1. As shown, the reduction of supply reduces consumer and producer surpluses. In economic terms, the total welfare in the before condition is $W_0 = CS_0 + PS_0$;

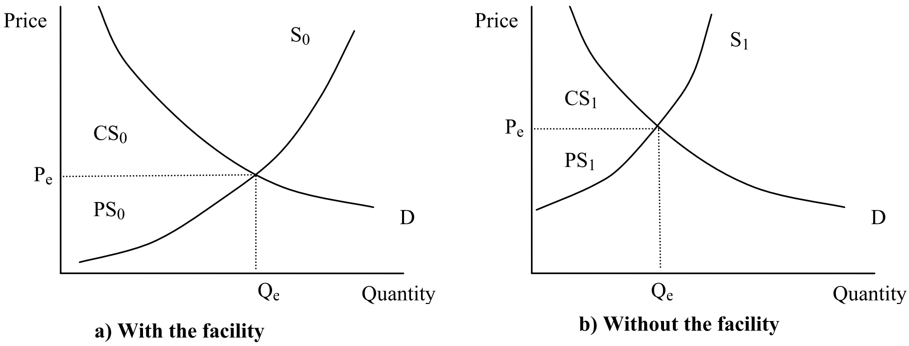


Figure 1. Consumer and Producer Surplus.

while the welfare for the after condition becomes $W_1 = CS_1 + PS_1$. The net economic impact of the change is: $B = W_1 - W_0$.

The reader must be aware that in economic analysis of transportation projects, the paradigm just described needs some significant adjustments. First, there is no formal supply function as in the production of goods; instead the equilibrium process is determined by the interplay between the demand function and the average costs and capacities perceived by the users. Second, since the equilibrium is determined by the costs perceived by the users, the externalities produced by the drivers are not taken into account. This leads to an imperfect market that, in the absence of optimal congestion pricing, would not achieve an efficient solution in which welfare is maximized (that occurs when the costs of externalities are internalized). Third, since transportation benefits and costs accrue over time, the analysis has to take into account how both producer and consumer surplus change with traffic conditions. For simplicity sake all these details are set aside for the time being.

In the context of a regional transportation planning agency, an estimate of the change in consumer surplus produced by a transportation project could be obtained with the assistance of a suitable regional transportation model. Running the model with and without the facility in question, would lead to two sets of values of generalized costs (C_{ij}) and equilibrium origin-destination flows (Q_{ij}). Assuming a linear relationship between the two equilibrium solutions, which leads to the rule of half, and summing across all origins and destinations (i, j), the change of consumer surplus could be estimated as:

$$\Delta CS = \sum_{i,j} \left[\left(C_{ij}^1 - C_{ij}^0 \right) Q_{ij}^1 + 1/2 \left(C_{ij}^1 - C_{ij}^0 \right) \left(Q_{ij}^0 - Q_{ij}^1 \right) \right] \qquad (3)$$

where the superscripts (0) refers to the initial condition and (1) the condition after.

Computation of equation (3) requires running the transportation model twice (with and without the project) for each time period of analysis, which is required to assess how changes in consumer surplus evolve over time, for the duration of the disruption. Although relatively simple to compute, the metric shown in equation (3), has a fundamental limitation: it does not take into account the likely income effects – associated with a major change in the transportation network – that are likely to take place. These changes in disposable income are bound to move upward (in the case of an income increase) or downward (in the case of an income reduction) the demand function which introduce an error in equation (3), in what is referred to as income effect (as opposed the substitution effect, i.e., movement along the same demand function, considered by Marshall). This phenomenon was first analyzed by Hicks (1943) who concluded that the solution was to isolate the income and substitution effects. This could be illustrated with the aid of Figure 2. (after Brent, 1997), which shows the indifference curves for two different situations: with (U_0) and without the facility in question (U_1). As shown, the utility of the initial condition is larger than the utility without the facility.

Figure 2 shows in the vertical axis a numeraire that represents all goods and services except the consumption of good X (transportation service through the facility). As shown, in the initial condition, the budget line (1) intercepts indifference curve U_0 at point A, leading to an optimal level of consumption equal Q_0. After removal of the facility, the likely increase in transportation costs would lead to an increase in the slope of the budget line (2), which pivoting around its intercept will become tangent to another indifference curve U_1 at B (leading to a consumption of Q_1). As outlined by Hicks (1943), in order to compensate the users for the loss of utility at the new prices, the users would have to receive an additional income equal to Δm_C (obtained by translating budget line 2, that represents the new prices, until it becomes tangent to the original indifference curve). (This increase represents the income effect.) Since the income effect has been isolated, the new point D (in the original indifference curve) represents what Hicks (1943) defined as compensated demand.

A second way to isolate the income effect is to do the analyses with the new indifference curve (U_1). Translating the budget line 1 until it is tangent to U_1 leads to point C. The difference in income between the original intercept and the intercept of the translated budget line two, i.e., Δm_E, represents the amount of income users would be willing to give up in order to keep the price at its original level. The resulting point c represents the equilibrated demand.

The lower part of Figure 2 shows the changes in consumer surplus for the original demand (P_0abP_1), for the compensated demand (P_0cbP_1) and the equilibrated demand (P_0cdP_1). These alternative measures of consumer surplus are usually referred to as: Marshallian (M), Compensating Variation (CV) and

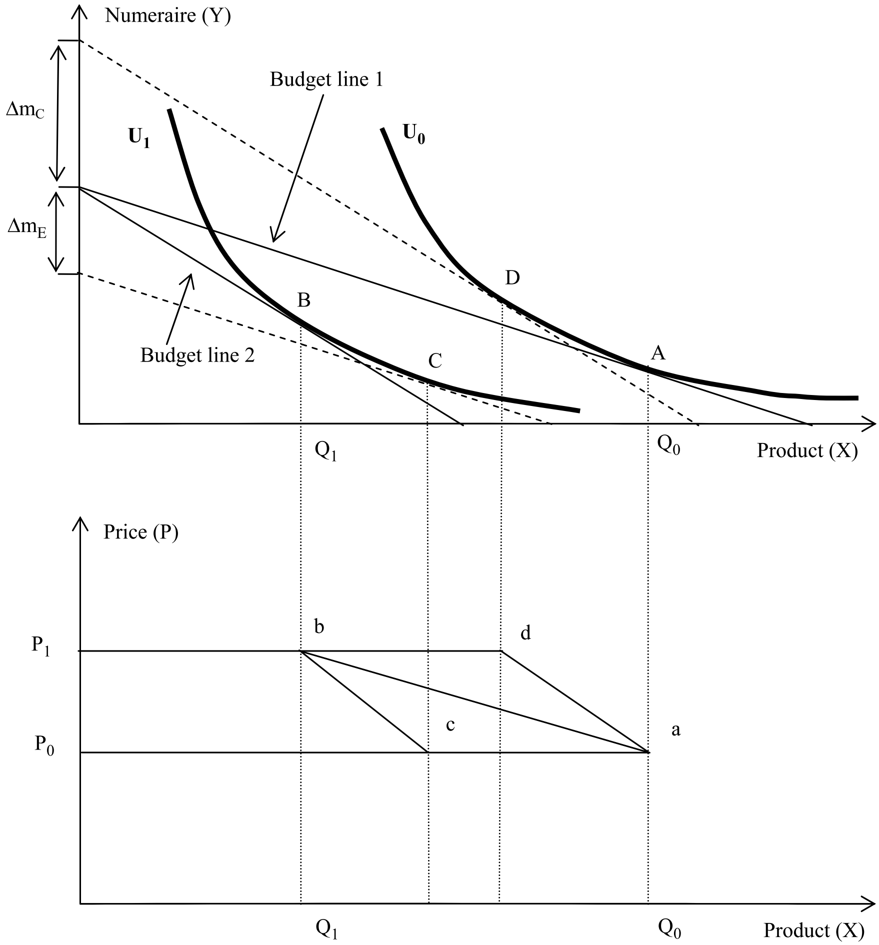


Figure 2. Hickian measures of consumer surplus.

Equilibriating Variation (EV). As shown, in the case of a loss of a facility, the CV will always be smaller or equal to M, while EV will always be larger an equal to M (the order is reversed in the case of an improvement, e.g., adding a transportation facility).

It is clear that the proper metric to use is the equilibriating variation (EV) because the issue at hand is one of disruption avoidance, in which the important question is how much each user is willing to pay to keep the status quo. Unfortunately, the computation of EV cannot be easily integrated to the standard transportation model because it requires specific knowledge about income effects. (The authors are not aware of the existence of any approximation formulas, such as the one by Willig (1976) that links M and CV, for EV and M.) In the absence of more appropriate measures, there is no choice but estimating the value of a facility on the basis of the Marshallian measure (M) which, as discussed before, is likely to overestimate the facility's economic value.

In spite of its intellectual appeal, the economic approach discussed in the previous section, that uses the Marshallian measure of consumer surplus, requires a relatively sophisticated use of the transportation models usually developed for Metropolitan Planning Organizations (MPOs). As shown before, the output of a transportation demand model would become the input of the economic analysis. This requires economic training of the kind not typically found at even the largest MPOs. The objective of the next section is to propose the use of alternative approximate techniques that may prove useful in providing approximate insights of the relative importance of individual facilities, without having to conduct complex modeling.

3. Approximation Techniques for Assessing Criticality in Transportation Networks

This section discusses three different models for modeling criticality in transportation networks. The first two models are approximate models which assess the components in the transportation network. The final model, a network interdiction model, identifies the different components which are to be protected without assessing the importance on individual components, or attempting to prioritize them.

The starting point for the discussion is equation (3) that shows how to compute the change in consumer surplus produced by a change in transportation supply. In its simplest form, the economic impact of a capacity change on the users is a function of both the change in demand and generalized costs. This suggests that a way to simplify the assessment process would be to use techniques that consider only one of these variables, either the change on network demand, or the change in transportation cost. These alternative metrics are discussed next.

3.1 Approximations Based on Traffic Flows

Probably, the simplest approximation of importance to the transportation facility is the one that relies on the number users of the facility in question. Using data provided by traffic counts has some practical advantages because traffic counts are: (1) readily available; (2) relatively accurate; (3) easy to update; and (4) routinely (and in some cases continuously) collected by transportation agencies. However, traffic counts are only remotely connected to the indicator of benefits shown in equation (3) which depends on origin-destination flows.

Nevertheless, traffic data do provide an indication of the intensity of use. In this context, defining t_{ij} as the traffic between two connected nodes i and j, one could compute a set of approximate indicators of criticality for both links and nodes, under the assumption that the criticality of a facility is proportional to its number of users. However, it shall be clear that this kind of metric does not attempt to compute the economic consequences of removing a link or a node from the transportation network (because these changes manifest themselves in the form of changes in generalized costs that these approaches fail to consider).

3.2 Approximations Based on Network Analysis: Network Science Models

The second approximate techniques to assess critical components accounts for network topology and measures the efficiency of the components based on the shortest path costs (as opposed to changes in the traffic level discussed in the previous section). These techniques consider the impact that removing a link or a node has on the level of connectivity of the network. These techniques are based on graph theory with link weights to represent network attributes. For example, the weight of each link could represent the distance, cost or a combination of the distance and the flow of material (people) between the nodes i and j. To find the critical components of G, the efficiency of the graph G must be defined first. The efficiency of G, as defined by Crucitti et al. (2003), is calculated on the basis of the shortest paths (d_{ij}), between any two nodes i(origin) and j(destination). The shortest path is computed using a label setting algorithm (see for instance, Ahuja et al., 1993). Assuming that the typical spatial interaction principles apply to the exchanges of goods or people along the network, the efficiency κ_{ij} of the flow between node i and j is inversely proportional to the shortest path distance d_{ij}, i.e., $\kappa_{ij} = \frac{1}{d_{ij}}$, $\forall i, j$. The *average efficiency* of the graph G can be defined as following:

$$\Xi(G) = \frac{\sum\limits_{\forall i \neq j \in G} \kappa_{ij}}{N(N-1)} = \frac{1}{N(N-1)} \sum_{\forall i \neq j \in G} \frac{1}{d_{ij}} \tag{4}$$

where $\Xi(G)$ denotes the efficiency of graph G. The above equation gives the value of $\Xi(G)$ in the range of $[0, \infty)$. To normalize this value we define the most efficient network in which there is a direct connectivity between nodes i and j, where flow moves through the network in the most efficient manner following the definition of Latora and Marchiori, (2004). Hence, $\Xi(G^{ideal})$ is defined as the most efficient network when all the nodes in a network are connected with all the other nodes. This is equal to:

$$\Xi(G^{ideal}) = \frac{1}{N(N-1)} \sum_{\forall i \neq j \in G} \frac{1}{x_{ij}} \quad (5)$$

where x_{ij} is the most efficient way of reaching j from i in a fully connected network. In this research efficiency is calculated for the transportation network by dividing it with $\Xi(G^{ideal})$ to get a value between[0, 1]. The harmonic mean is used to calculate the efficiency in (4) and (5) as the harmonic mean was found to be a better average estimator than arithmetic mean to characterize the mean flow of information in communication networks (Latora and Marchiori, 2003).

With the overall efficiency of a graph G in place, the value of each node i to the overall system can be calculated by computing the efficiency of the graph G with and without the node in the graph. A new graph is defined by removing the facility under analysis, which leads to a new graph G^{-i}, which does not include facility i. The importance of facility i is calculated as:

$$\Im(i) = \Xi(G) - \Xi(G^{-i}), \quad \forall i \in V(G) \quad (6)$$

where, $\Im(i)$ indicates the value of link i to the overall network and $\Xi(G^{-i})$ represents the efficiency of the network by deactivating the node i in graph G. Based on the magnitude of $\Im(i)$, the critical components of the network are identified as the ones that have the *highest value* of $\Im(i)$. This analysis can be used to identify the critical components of the network.

The main algorithm used in the above methodology is the shortest path (SP) algorithm. This is a well researched problem and can be solved very efficiently. The worst case complexity of Dijkstra's algorithm is known to be $O(n^2)$, where n is the number of nodes. However, the complexity of identifying critical components is greater, as this operation is performed n times, for each facility in the network with and without it. Therefore, the total SP calculation in the base network is $O(n^3)$. The same complexity holds for calculating the most efficient network. The importance of each node i requires the above operation of calculating the one to all SP's |V| times for each node. Overall, the algorithm requires $O(n^4)$ worst case time for running this algorithm. Hence, we conclude that the worst case complexity of the above procedure is $O(n^4)$. Since, the worst case complexity is in polynomial time, this method can be used efficiently to identify critical components for large scale networks.

3.3 Network Interdiction Models

This section discusses a special kind of models that, although not directly concerned with assessing how critical a component of the network is, provide insight into the strategic behavior of an adversary interested in maximizing disruptions to the transportation network, and those interested in interdicting these agents from causing network damage.

In a simplistic fashion, three agents are at play. The first one is the agent interested in maximizing network security (broadly defined), the interdictor. In reality, this "agent" may in fact represent a collection of agencies, or individuals. However, for the purposes of this paper, it is assumed that their coordinated actions could be represented by the actions of a (super) agent. The second agent represents the adversary that threatens to disrupt the transportation network, e.g., terrorists trying to blow up a bridge. Another type of a disruptive agent are the natural forces (e.g., hurricanes, earthquakes) that tends to have a random and significant impact on the transportation network, as demonstrated by hurricane Katrina. For purposes of discussion, these three agents are referred to as: interdictor, adversary and nature. It shall be clear that although both adversary and nature can and do produce significant network disruptions, only adversary and interdictor engage each other in the tactical and strategic interactions studied by game theory (on which the discussion in this section is based on). However, nature could be incorporated into this type of models by replacing the original objective function (disruption maximization), with probability functions that reflect the likelihood of damage to individual components by a natural event.

In this context, it is obvious that another distinction must be made between two important cases that differ based on the information that the interdictor and adversary know about each others' strategies. The first one considers the case in which interdictor takes actions to reduce the vulnerability of a transportation facility, and attackers are aware of these strategies. For instance, road inspections at critical transportation facilities are likely to be known to potential attackers. This situation could be interpreted as a Stackelberg game with the interdictor as the leader of the game, and adversary as the follower. Some type of network interdiction models are leader-follower games where the two players are in a warlike conflict (Wood, 1993). The follower operates a network in order to optimize a particular objective function such as moving container shipments through a given network as quickly as possible. The leader attempts to limit the follower's objective by interdicting arcs, for example, by attacking arcs to destroy them, either to completely eliminate the movement or reduce the capacity.

The second case is related to a situation in which neither interdictor nor the adversary are unaware of the strategies of each other, although they have information on the set of all available strategies. An example of this would be a situation in which interdictor tries to minimize disruptions in the network by conducting random inspections at key locations; while the adversary is doing the opposite, i.e., trying to maximize network disruptions. This is in essence a non-cooperative game. Some formulations of network interdiction models (e.g., Salmeron et al., 2004) are equivalent to a non-cooperative game with a Nash equilibrium solution, as shown in P1.

$$\textbf{(P1)}\quad \max_{\delta \in \Delta} \min_{p} \mathbf{c}^{\mathbf{T}}\mathbf{p} \tag{7}$$

$$s.t.$$

$$\mathbf{g}(\mathbf{p}, \delta) \leq b \tag{8}$$

$$\mathbf{p} \geq 0 \tag{9}$$

An interdiction plan is described by a binary vectorδ, whose kth entry δ_k is 1 if component k of the system is attacked and is zero otherwise. For a given interdiction plan, the inner problem is an optimal traffic-flow problem that minimizes either the cost of the entire system (System Optimal Assignment) or each user's cost (User Equilibrium). This is denoted by $\mathbf{c}^{\mathbf{T}}\mathbf{p}$. The vector $\mathbf{p}$ denotes the traffic flow in the transportation network in consideration and $\mathbf{c}$ represents the generalized non-linear costs on each link of the network. The outer maximization chooses the most disruptive plan from the external agents viewpoint, $\delta \in \Delta$ where Δ represents the discrete set of attacks that an external agent will be able to carry out. The set of constraints corresponds to a set of functions that involve the flow balancing constraints. Other additional constraints would depend on the specific transportation network structure (e.g., multiple modes, user groups) in consideration. Most of the interdiction problems involve a binary decision variable with a deterministic outcome. The *k-most vital arcs problem* (Corely and Shaw, 1982) is usually a special case of the interdiction problems, in which the interdictor seeks to destroy exactly k arcs to interdict the network most effectively. The k-most vital arcs problem being NP-complete (Ball et al., 1989), it follows that most of the interdiction problems are NP-complete and hence computationally difficult to solve. However, many approaches like Benders decomposition, super valid inequalities (Israeli and Wood, 2002) and new bi-level algorithms (Salmeron et al., 2004) have been developed to solve real size networks.

The models discussed in this section consider the strategic behavior of the adversary who wishes to infiltrate a road network to send materiel or cause damage and the interdictor wishes to minimize the damage by interdicting arcs in the network. Different variations of this problem could be studied under different assumptions such as budget constraint for the interdictor; where he/she cannot spend more than a pre-specified amount on the interdiction plan. Further extensions include accounting for stochasticties in arc capacities and network demands leading to more difficult formulations and solution approaches. Some of these models have been demonstrated for defense related problems such as nuclear smuggling problems (Morton and Pan, 2004). Note, that the final result from these models gives the network arcs that are interdicted rather than any prioritization scheme for the arcs or nodes. This differentiates this modeling approach from the ones proposed in 3.1 and 3.2.

4. Conclusions

The paper discusses alternative modeling techniques that could be used to assess the criticality of transportation facilities. The paper starts with a comprehensive discussion of the economic techniques to assess the economic value of transportation facilities. The authors conclude that the traditional (Marshallian) measure of consumer surplus is likely to overestimate the economic value of a facility, because of the lack of consideration of income effects. After discussing several metrics to assess the consumer surplus, the authors conclude that the most appropriate metric is what Hicks (1943) defined as "equilibriating variation." This is because this is the metric that provides an estimate of the economic value of maintaining the status quo conditions in the network.

The paper then discusses approximate techniques to assess criticality of transportation facilities. Two set of approximation techniques, based on traffic flows and shortest path distances, are discussed. The technique that focuses on network changes was further studied in connection with network science concepts. A final group of techniques, based on the notion of network interdiction models are also analyzed.

A contribution of this paper is in proposing a model based on complex network science for identifying critical components in a transportation network. This can be applied to identify critical components in a transportation network. This methodology has potential to be applied for identifying critical components in large transportation networks when the interactions can be estimated.

By systematically studying the interventions in this manner, organizational responsibilities, technology investments and mitigation efforts can be identified and preparedness for external threats be accommodated in a holistic manner. The external threats to transportation systems are critical and it is imperative to interdict and minimize them. However, it is also important to strike a balance between the vital economic function of the transportation system (i.e., to ensure the seamless movement of goods and people) and the security interventions. This raises challenging issues which are yet to be resolved in securing critical transportation networks against hostile threats. As mentioned before, there are key challenges with organizational responsibilities, funding, reliable information about passenger/goods movements, and tradeoffs between trade facilitation versus transportation security. Modeling these offers a unique perspective in gaining insights into decision making which otherwise would be conjecture and ad-hoc. Securing transportation systems is a priority and to be effective it must be organizationally, strategically and economically sustainable.

References

Ahuja, R.K, T.L. Magnanti and J.B. Orlin. *Network Flows*. Prentice-Hall, Englewood Cliffs, NJ, (1993).

Ball, M.O., Golden, B.L. and Vohra, R.V. "Finding the most vital arcs in a network." Oper Res Lett 8 (1989) 73–76.

Brent, R.B. "Applied Cost-Benefit Analysis" Edward Elgar, publisher, (1997) ISBN 1-85898-338-X

Corely, H.W. and Shaw, D.Y. "Most vital links and nodes in weighted networks." Oper Res Lett 1 (1982) 157–160.

Curcitti, P., Latora, V., Marchiori, M., and Rapisarda, A. "Efficiency of scale-free networks: error and attack tolerance". *Physica A* 320 (2003) 622–642.

Dupuit, J. "On the Measurement of the Utility of Public Works" reprinted in International Economic Papers, No. 2, London: Macmillan.

Elliott, E and Kiel, L.D. "A complex systems approach for developing public policy towards terrorism: an agent based approach." Chaos, Solitons and Fractals 20 (2004) 63–68.

Epstein, J.M., R. Axtell *Growing Artificial Societies: Social Science from the Bottom Up*, The MIT Press, Cambridge, MA, 1996.

Harrald, J.R., H.W. Stephens, and J. Rene vanDorp (2004) "A Framework for Sustainable Port Security", *Journal of Homeland Security and Emergency Management*. 1 (2), Article 12.

Harvey, D., M. Reed "Social Science as the Study of Complex Systems," In *Chaos theory in the Social Sciences: Foundations and Applications*, L. D. Kiel, E. Elliot (Ed.), The University of Michigan Press, An Arbor, 1997, pp. 295–323.

Hecker, JayEtta Z. "*Port Security: Nation faces formidable challenges in making new initiatives successful*" U.S. General Accounting Office, August 1, 2002. Washington D.C.

Hicks, J.R. "The Four Consumers' Surpluses," Review of Economics Studies, Vol. 11, pp. 31–41 (1943).

Israeli, E and Wood, R.K. " Shortest-Path Network Interdiction." Networks 40(2), (2002) 97–111.

Latora, V. and Marchiori, M. "How the science of complex networks can help develop strategies against terrorism". *Chaos, Solitons and Fractals* 20(2004) 69–75.

Maritime Transportation Security Act, 2002. (http://www.uscg.mil/hq/g-m/mp/pdf/MTSA.pdf) Accessed on May 23rd, 2004.

Marshall, A. "Principles of Economics," (8th Ed.) London, Macmillan (1924)

Morton, D. and Pan, F. "Stochastic Network Interdiction of Nuclear Material Smuggling", Presented at the Xth International Conference on Stochastic Programming, Tuscon, AZ, 2004.

Resnick, M. *Turtles, Termites, and Traffic Jams*, The MIT Press, Cambridge, MA, 1994.

Salmeron, J., Wood, R.K. and Baldick, R. "Analysis of Electric Grid Security under terrorist threat." *IEEE Transactions on Power Systems* 19(2) May 2004.

Simon, H.A. “Prediction and Prescription in Systems Modeling.” *Oper Res.* 38 (1), 1990, 7–14.
Smith, T. ”Nonlinear Dynamics and the Micro-Macro Bridge,” In *Chaos, Complexity and Sociology*, R. Eve, S. Horsfall, M. Lee (Ed.), Sage Publications, Thousand Oaks, 1997, pp. 52–78.
Wood, R.K. “Deterministic Network Interdiction,” *Mathematical and Computer Modelling* 17, 1993, pp. 1–18.
Willig, R.D. “Consumer Surplus Without Apology,” American Economic Review, Vol. 66, pp. 589–597 (1976).

Chapter 10

A Simulation-Based Dynamic Intermodal Network Equilibrium Algorithm

Elaine Chang[1] and Athanasios Ziliaskopoulos[2]

[1]Department of Civil and Environmental Engineering, University of South Florida, 4202 E. Fowler Ave. ENB 118, Tampa, FL 33620, E-mail: echang@eng.usf.edu

[2]Department of Industrial and Mechanincal Engineering, University of Thessaly, Pedion Areos, Volos, 38222 GREECE, E-mail: ziliasko@yahoo.com

Abstract This paper introduces a Variational Inequality (VI) formulation for the time-dependent combined mode split and traffic assignment problem. Travel costs are represented by generalized cost functions and mode choices are deterministically obtained based on assignment to intermodal least cost paths without accounting for possible randomness in travelers' choices. The intermodal user equilibrium (IUE) is estimated using an inner approximation (IA) algorithm that results in a nonlinear program with linear constraints. The algorithm converges assuming continuous and monotonic path travel cost functions. The paths on multimodal networks are computed with an intermodal optimum path algorithm; a cell transmission-based simulator, enhanced to account for both automobile and transit vehicles, is used to estimate the path travel costs. A heuristic search approach is proposed and implemented in the VISTA simulation-based framework. Computational results are presented on example networks to test convergence and equilibrium.

Keywords: user equilibrium; traffic assignment; variational inequalities; simulation

1. Introduction

In decisions regarding construction of transportation infrastructure, adoption of transportation policies and implementation of transportation technologies, mathematical models are often used to predict or estimate the impacts of the alternatives under consideration. Mathematical models are performed before or in place of costly field tests and full-blown implementations. The challenge in transportation modeling is to develop models that correctly represent traveler behavior, while remaining mathematically tractable. Travel behavior is extremely complex to model as each trip includes different types of choices, including, but not limited to, destination choice, mode choice and route choice. Traditionally, urban transportation planning has treated each type of choice in separate models, thus ignoring the interdependence of these choices within a trip. Further, the state of the practice continues to model travel behavior in a static realm, wherein travel conditions are averaged over the time period being modeled, typically one or more hours representing a peak period or off-peak period. As a result, time-varying travel conditions, such as congestion, queuing, and transfer delays, are ignored.

This paper discusses a network-based approach to estimating intermodal route choice, accounting for time-varying realities in urban transportation networks, such as congestion, queuing and transfer waiting times. The literature shows that several combined mode and route choice models have been proposed for the static problem; however, static combined models do not provide a realistic representation of time-varying traffic conditions. Moreover, while much research has been done in the field of dynamic traffic assignment (DTA) to capture time-varying traffic conditions, much of this work has focused on automobile movements and automobile route choices, under the assumption that mode choice is relatively unaffected by the minute-to-minute dynamics of traffic conditions. As information technology matures, mode choice will likely become a real-time decision, instead of a long term planning decision separate from the route choice.

The proposed intermodal combined mode and route choice model predicts person trips over a multi-modal network, such that mode, route and transfer choices are modeled as simultaneous decisions. As such, changes in congestion resulting from changes in mode choices are captured. Further, ridership and mode share, which are central issues in transit policy evaluation, may be directly observed in the model. The intermodal route choices are determined using a decomposition algorithm, which iterates among traffic simulation, route cost calculation and allocation of flows (trip assignment) to minimize a gap function corresponding to the variational inequality formulation of the dynamic intermodal equilibrium problem. An implementation of the algorithm is proposed that relies on a cell transmission-based traffic simulator to maintain feasibility of the vehicle movements, and to evaluate the path travel times and transit transfer times. The vehicle simulation model captures queuing, signal delays and bus movements. Further, intermodal paths are generated using an algorithm that calculates time-varying least-cost paths. The algorithm is shown to converge for cases with continuous, monotonic cost functions, and computational experiments are performed to explore convergence properties for less ideal conditions. For example, tests with varying levels of demand, signal timing and transfer waiting times are tested to explore their effects on monotonicity of the cost functions and resulting gap function, as well as on convergence of the algorithm.

2. Background

Mode split and traffic assignment models have traditionally been estimated in separate steps; however, these decisions are, in fact, interdependent. Over the past few decades, several approaches to combining these steps have been proposed in an effort to capture the interaction between mode and route choices. Early research in this area saw the development of many static combined mode split and traffic assignment models; however, static combined models do not provide a realistic representation of time-varying traffic conditions. Later, advancements in dynamic traffic assignment (DTA) improved the realism with which time-varying traffic conditions are captured, but the bulk of work in DTA has focused on the propagation and route choices of automobiles, taking automobile trip tables as an exogenous input, and thus assuming that mode choice is relatively unaffected by the minute-to-minute dynamics of traffic conditions. However, as information technology matures, it is expected that mode split will become a real-time decision, instead of a long-term planning one separate from the route choice. As such, a model's ability to capture both mode and route choices will enhance the realism of the model, since changes in congestion resulting from different mode choices can be captured. Further, ridership and mode share, which are central issues in transit policy evaluation, can be directly observed in a combined model.

Early approaches to modeling the combined static mode split and assignment problem were proposed by Florian (1977), Florian and Nguyen (1978), Abdulaal and LeBlanc (1979). These models consisted of mathematical programming (MP) formulations of the network equilibrium with logit functions to capture the mode split. Aashtiani (1979) formulated the traffic equilibrium problem as a non-linear complementarity problem, which could also be applied to multi-class, multimodal and destination choice user equilibrium problems. Later, Dafermos (1982) and Florian and Spiess (1983) presented variational inequality (VI) formulations of the problem. Each of these formulations assigned trips only to pure modes, without considering mid-trip mode transfers.

Fernandez et al. (1994) presented a MP model that allowed for mid-trip mode transfers by having the more meaningful modal combinations defined as new modes in the mode choice set. Boile, Spasovic and Bladikas (1995) presented a more flexible intermodal model with pure auto trips on one branch and rail and intermodal trips on another branch. The route and transfer flows are then determined based on a user optimal assignment, such that mode transfer options are not required to be explicitly defined.

More recently, Abrahamsson and Lundqvist (2003) proposed combined equilibrium models that solve the destination and mode choices simultaneously, and then iterate that result with the route choice. Wu and Lam (2003) presented a VI formulation of the combined mode split and stochastic assignment problem to estimate intermodal pedestrian and transit trips under congested conditions, but did not include automobile trips in their model. De Cea, Fernandez and Dekock (2003) proposed a VI formulation of the combined mode split and assignment problem to be solved using a diagonalization algorithm within the ESTRAUS modeling framework.

These static approaches are limited by their inability to capture the time-dependent network conditions. Specifically, link travel times, which are calculated using link performance functions, are assumed to be static through the model period, such that congestion and queuing effects are ignored. Further, static assignment models do not capture the exact arrival and departure times of buses, so waiting times are estimated based on transit schedule frequency, if they are considered at all.

The combined mode and route choice problem has only recently been addressed in a time-dependent context. Two simulation-based models, DYNASMART (Mahmassani, Peeta and Ziliaskopoulos (1993)) and TRANSIMS (Los Alamos National Laboratories (2002), have the capability of modeling intermodal trips. DYNASMART was originally developed to solve the automobile-only DTA problem. Its simulator capabilities were extended by Abdelghany and Mahmassani (2001) to capture bus movements, such that buses are tracked along their routes their arrival times at each stop are recorded. Further, link capacity and flow rate are adjusted for the duration that a bus is stopped on a given link. Intermodal paths are generated by a multi-objective routing algorithm. Vehicle trips are simulated such that if a trip includes a car portion for the whole trip or part of it, then a car is generated and moved into the network. Transit vehicles are generated according to a pre-determined schedule and follow pre-determined routes. As with the single-mode DYNASMART model, the assignment pattern is determined using a Method of Successive Averages approach.

The TRANSIMS model is an activity-based model with modules that estimate such trip aspects as activities, destinations, modes and routes. A traffic simulator models vehicle movements, based on which travel times may be determined. The trip assignment pattern is determined by an iterative heuristic algorithm, for which no proof of convergence has been provided.

More recently, Benjamins, Lindveld and van Nes (2002) and Carlier et al. (2003) proposed a model architecture for multimodal transport modeling. They recasted the combined mode and route choice problem as a pure route choice problem within a multimodal network, and formulated the problem as a VI that assumes a stochastic UE. The solution algorithm iterates through path-finding and route-choice, which are modeled using random utility theory in separate modules. Multi-user class assignment is used to accommodate the heterogeneity of choice preferences. No proof of convergence or correctness were provided for the solution algorithm.

This paper presents a VI formulation of the time-dependent combined mode split and traffic assignment problem. Travel costs are represented by generalized cost functions, and mode choices are deterministically obtained based on assignment to intermodal least cost paths, and no stochastic, logit-type relations are used. The user equilibrium intermodal trip assignment is estimated using an inner approximation (IA) algorithm, which searches the feasible space for the assignment that minimizes the equilibrium gap function. The algorithm is proven to converge for cases with continuous, monotonic cost functions; however, since cost functions in intermodal networks may not be continuous and monotonic a heuristic search approach is proposed, and computational results are presented for a test network.

3. Model Formulation of Intermodal Trip Equilibrium

For the intermodal assignment problem, we consider a multimodal network *G(M,V,A,T)* where M is the set of available modes, V is the set of nodes, A the set of arcs and *[0,T]* the assignment period. The network may be considered a series of modal subnetworks, such that each modal subnetwork contains nodes and links, and transfers between modes are limited to transfer links, which connect modal subnetworks.

Next, let d^t_{rsb} be the number of trips from node r and to node $s(r, s \in V)$ at time $t \in [0, T]$ and generated by travelers of behavior type b. The behavior type is associated with a set of cost parameters within a generalized cost function to define travelers' travel preferences.

Further, let P be the set of all spatiotemporal paths from all origins to all destinations, i.e. $P = \{p^1, p^2, \ldots p^\pi\}$. Each path p^k, $1 \leqslant k \leqslant \pi$ belongs to a set *P(r,s,t,b)*, which contains all paths defined for travelers of behavior type b departing at time t in $[0, T]$ from node r to node $s(r, s \in \ V)$.

As with the single-mode version of the algorithm, we denote with ξ^{p^k} the number of travelers choosing to follow intermodal path p^k(Ξ in vector notation), and $\psi^{p^k}(\Xi)$, the travel cost on path $p^k(\Psi(\Xi)$ in vector notation). The intermodal travel cost includes different kinds of travel and transfer costs, and is thus calculated based with a generalized cost function. The function includes weighted costs of fixed and time dependent travel and transfer costs.

Further, the cost parameters or weights may be defined differently for behavior types to reflect differences in individual travel choice preferences.

The demand relationships $\sum_{p^k \in P(r,s,t,b)} \xi^{p^k} = d^t_{rs}$ form a closed, bounded, convex space $D \subset R^{\pi}$. As such, any assignment Ξ in D is feasible, given that the traveler propagation law adopted allows all travelers to complete their trips within time T.

The Wardrop equilibrium is reached when no traveler has a less costly alternative route. Further, it is assumed that a travelers' selection of an alternative path is a unilateral decision based on the current traffic conditions. The Wardrop user equilibrium intermodal path assignment can be defined mathematically as Ξ^* in D where

$$\psi^{p^x}(\Xi^*) > \psi^{p^y}(\Xi^*) \quad \text{implies that } \xi^{p^x} = 0 \; \forall p^x, p^y \in P(r,s,t,b) \; \forall r,s,t,b$$

In other words, at the equilibrium path assignment, any path that is more costly than the minimum cost path remains unused. Alternatively, it can be stated that

$$\Psi\left(\Xi^*\right)^T \Xi > \Psi\left(\Xi^*\right)^T \Xi^* \; \forall \Xi \in \mathrm{D}$$

This formulation is less strict than the previous one as it allows some travelers to choose more costly routes, as long as the total route cost is reduced; however, the equilibrium point coincides with the Wardrop equilibrium.

Suppose we have a solution $\Xi' \in \mathrm{D}$ such that

$$\psi^{p^x}\left(\Xi'\right) > \psi^{p^y}\left(\Xi'\right) \quad \text{for some } p^x, p^y \in P(r,s,t,b) \text{ and } r,s,t,b$$

but that $\xi^{p^x} > 0$ for some p^x, where $p^x \neq p^y \xi^{\text{rstbp'}} > 0$. This suggests that switching vehicles along $\mathrm{p^x}$ to the cheaper route $\mathrm{p^y}$ will reduce the total cost by

$$\psi^{p^x}\left(\Xi'\right)\xi^{p^x} - \psi^{p^y}\left(\Xi'\right)\xi^{p^y} > 0$$

Therefore, if the resulting path assignment is Ξ'', then

$$\Psi\left(\Xi'\right)^T \Xi'' < \Psi\left(\Xi'\right)^T \Xi' \; \forall \Xi'' \in D$$

As such, we have shown that

$$\psi^{p^x}(\Xi^*) > \psi^{p^y}(\Xi^*) \quad \text{implies that } \xi^{p^x} = 0 \; \forall p^x, p^y \in P(r,s,t,b) \; \forall r,s,t,b$$

if and only if

$$\Psi\left(\Xi^*\right)^T \Xi < \Psi\left(\Xi^*\right)^T \Xi^* \; \forall \Xi \in D$$

Therefore, an equilibrium solution, $\Xi*$, exists where

$$\Psi\left(\Xi^{*}\right)^{T}\left(\Xi-\Xi^{*}\right)\geq 0\;\forall\Xi\in D \tag{1}$$

Relationship (1) is a variational inequality (VI) formulation of the dynamic user equilibrium assignment problem. Smith (1979) and Dafermos (1980) showed for the static traffic assignment problem, the VI formulation with space D being compact and $\Psi(\Xi)$ assumed to be monotonic and continuous, there exists a unique solution.

For the dynamic intermodal trip assignment problem of the same form, it is reasonable to assume that the same proofs hold given the same assumptions. Specifically, based on Smith's and Dafermos' work, we claim that a solution Ξ^* to the VI formulation (1) exists where the space D is compact and $\Psi(\Xi)$ is assumed to be monotonic and continuous. In fact, generalized cost functions for intermodal problems are rarely continuous, due to traffic signal delays and congestion, as well as transit fares and intermodal transfer times. As a result, the existence of an equilibrium point is not guaranteed in most realistic problems. Further, due to temporal interactions of travelers assigned earlier with those assigned later, monotonicity may not hold. Lack of monotonicity precludes proof of solution uniqueness, as well as algorithm convergence. However, in practice the assumption of monotonicity is reasonable, since path costs tend to be dominated by number of travelers on that path, rather than those on other paths, and thus that costs do not decrease with number of travelers on a path. In addition, while in reality the cost function may be discontinuous, and thus an exact solution may not exist, it is still possible to try to approach the equilibrium condition where nobody can switch to a lower cost path.

At this point we have not proposed a specific cost function, or even a form of a cost function, because the VI formulation applies with any generalized cost function, keeping in mind that continuity and monotonicity of cost functions ensure existence and uniqueness of a solution.

In the next section, a solution algorithm is proposed, and convergence of the algorithm is proven for the case where the cost function is assumed to be continuous and monotonic. However, as explained, since cost functions are typically neither continuous nor monotonic in reality, the algorithm is recognized to operate as a heuristic, and alternative implementations that appear to improve convergence in practice are discussed.

4. Inner Approximation Solution Algorithm

The algorithm proposed in this section estimates the equilibrium path assignment using inner approximation methods, and has thus been named the

Inner Approximation (IA) dynamic user equilibrium algorithm. Conventionally, vehicle assignment has been performed using the method of successive averages (MSA), which assigns trips equally among the set of past solutions. In contrast, rather than assuming that all previous solutions contribute equally to the final equilibrium, the IA algorithm searches the feasible set of path assignments for the assignment that minimizes an equilibrium gap function.

The name of the algorithm, Inner Approximation, was selected because the procedure searches within a subspace that is defined by a set or subset of the extreme points of the feasible space. The approach is similar to simplicial decomposition, but differs in the descent direction used at each iteration. Specifically, with simplicial decomposition the descent direction is the extreme direction that results in the greatest improvement in the value of the gap function, whereas with inner approximation, the descent direction is an average of the extreme directions that violate the equilibrium condition.

An inner approximation approach is proposed to solve the intermodal DUE problem, where the travel time cost function is assumed to be continuous and monotonic. The approach minimizes a gap function defined as shown in Equation (2).

$$V_{Smith}(\Xi) = \sum_{i=1}^{N} \max^2 \left\{0, -\Psi(\Xi)^T (P_i - \Xi)\right\} \tag{2}$$

where P_i is an extreme point of the convex hull D, and N is the number of extreme points defined. This gap function was defined based on a gap function originally proposed by Smith (1983) for the static user equilibrium problem, and abides by the definition of a gap function, namely

$$\begin{aligned} &\text{i.} \quad V(\Xi) = 0 \quad \forall \Xi \in \Omega \\ &\text{ii.} \quad V(\Xi) \geq 0 \quad \forall \Xi \in D \end{aligned} \tag{3}$$

The proposed IA algorithm then iteratively selects a descent direction of V_{Smith} within the assignment space D, and a step length that minimize the gap function in that direction. The algorithm, shown in Figure 1, terminates when the value of the gap function is $V(\Xi) = 0$. The IA algorithm's search procedure may be implemented using any vehicle and person propagation relationships, and is presented here with path costs estimated based on simulation of travel conditions.

To find the minimum point of (2), a descent direction and step length are selected, a new solution is found and the search procedure is repeated. The descent direction is composed of a weighted average of the extreme directions that deviate from equilibrium, as shown in Equation (4). This direction is

Step 0. Initialize
Set $n = 0$.
Set link travel time to free flow.
Calculate least cost paths for each rstb.
Set Ξ_0 to all-or-nothing assignment of d^t_{rsb} to least cost path for $rstb$.
Simulate travel conditions with assignment Ξ_0.
Update link travel times.

Step 1. Choose new solution
Choose descent direction Δ_n.
Select step length $\lambda_n = \text{argmin}_\lambda(V_{Smith}(\Xi_n + \lambda\Delta_n))$ using golden section search.
Assign demand to $\Xi_{n+1} = \Xi_n + \lambda_n\Delta_n$.

Step 2. Update costs
Simulate traffic conditions with assignment Ξ_n.
Update link and path travel times.

Step 3. Check for convergence
If $V_{Smith}(\Xi_{n+1}) \geqslant 0$
  Set $n = n + 1$.
  Return to Step 1.
Else
  Terminate.

Figure 1. The IA algorithm.

labeled Δ_{Smith}, since it was originally proposed by Smith (1983) for application to the static assignment problem.

$$\Delta_{Smith\ n} = \frac{\sum_{i=1}^{N} \max\left(0, -\Psi\left(\Xi_n\right)^T \left(P_i - \Xi_n\right)\right) \cdot \left(P_i - \Xi_n\right)}{\sum_{i=1}^{N} \max\left(0, -\Psi\left(\Xi_n\right)^T \left(P_i - \Xi_n\right)\right)} \tag{4}$$

The new solution thus becomes

$$\Xi_{n+1} = \Xi_n + \lambda\Delta_{Smith\ n}$$

where n is the iteration number, $\Delta_{Smith\ n}$ is the descent direction for iteration n, and λ is the step length ($0 \leqslant \lambda \leqslant 1$). Smith (1983) showed that the direction Δ_{Smith} guarantees an improvement of

$$\Delta V = \nabla V(\Xi) \cdot \Delta\Xi \leq -2V(\Xi)$$

in the gap value with each step. While Smith's proof was presented for application to the static assignment problem, the logic applies more generally to VI problems of the form (1) under the assumption that the cost function is continuous and monotonic. As such, the proof of convergence using also applies to the intermodal dynamic trip equilibrium problem and the proposed IA algorithm.

In cases where the assumptions of continuous and monotonic cost functions do not hold, the algorithm is considered a heuristic, since the existence of a solution is not guaranteed when the cost function is not continuous. Further, a non-monotonic cost function may result in the search procedure terminating at a local minimum of the gap function without reaching a true equilibrium point. As in the single-mode case, since it is known that the global minimum of the gap function must be 0, and that this global minimum corresponds to an equilibrium path assignment, the gap value provides a convenient measure of the deviation of any given solution from the equilibrium condition.

For problems where the assumptions of continuity and monotonicity do not hold, the convergence of the heuristic algorithm with different search strategies is proposed as an implementation alternative to Smith's weighted average of the extreme points. This implementation alternative is described in the next section.

5. Heuristic Implementation

The proposed heuristic version of the IA algorithm includes three different search phases, each similar to the procedure outlined in Figure 1, but differing in their gap functions and descent directions. Based on numerical observations, the gap functions and descent directions of each phase are proposed for different stages of the search for the equilibrium solution.

The first phase of the IA algorithm uses an extreme direction search approach to minimize the gap function,

$$V_{extreme}(\Xi) = \sum_{i=1}^{N} \max\left\{0, -\Psi(\Xi)^T (P_i - \Xi)\right\} \tag{5}$$

where P_i is an extreme point of the convex hull D, and N is the number of extreme points defined. Each extreme point P_i represents an all-or-nothing path assignment, and any point in D, including the equilibrium solutions, $\Xi^* \in \Omega \subseteq D$, can be defined as a convex combination of the extreme points P_i.

Equation (5) is consistent with the definition of a gap function (3), so if the cost functions are reasonably approximated as continuous, monotonic functions, then the global minimum of the gap function corresponds to an equilibrium path assignment.

To find the minimum point of (5), a numerical search approach is used, such that a descent direction (in the assignment space D) and step length are selected, a new solution is found and the search procedure is repeated. In Phase 1, the algorithm selects, as the descent direction, an extreme point composed of an all-or-nothing assignment to the shortest paths. The new solution thus becomes

$$\Xi_{n+1} = (1-\lambda)\,\Xi_n + \lambda P_n$$

where n is the iteration number, P_n is the extreme point descent direction selected for iteration n, and λ is the step length ($0 \leqslant \lambda \leqslant 1$).

Upon termination of Phase 1, the algorithm continues with Phase 2, which is the weighted average approach described in described in Section 3. More specifically, during this phase the algorithm minimizes the gap function

$$V_{Smith}(\Xi) = \sum_{i=1}^{N} \max^2 \left\{0, -\Psi(\Xi)^T (P_i - \Xi)\right\}$$

by searching in the direction

$$\Delta_{Smith\ n} = \frac{\sum_{i=1}^{N} \max\left(0, -\Psi(\Xi_n)^T (P_i - \Xi_n)\right) \cdot (P_i - \Xi_n)}{\sum_{i=1}^{N} \max\left(0, -\Psi(\Xi_n)^T (P_i - \Xi_n)\right)}$$

The new solution thus becomes

$$\Xi_{n+1} = \Xi_n + \lambda \Delta_{Smith\ n}$$

where n is the iteration number, $\Delta_{Smith\ n}$ is the descent direction for iteration n, and λ is the step length ($0 \leqslant \lambda \leqslant 1$).

The gap function used in the third phase, based on the cost gap formulation of the equilibrium, is shown in Equation (6).

$$V_{cost}(\Xi) = \sum_{r \in V} \sum_{s \in V} \sum_{t \in [0,T]} \sum_{p \in P(r,s,t)} \xi^p \left(\psi^p(\Xi) - \psi^{p_{\min rst}}(\Xi)\right) \tag{6}$$

Equation (6) is consistent with the definition of a gap function (3), so if the cost functions are reasonably approximated as continuous, monotonic functions, then the global minimum of the gap function corresponds to an equilibrium path assignment.

In this phase, the descent direction is X_n, where each element of X_n is set to the values

$$x^{p^x n} = \begin{cases} \xi^{p^x n} & \text{if } p \text{ is neither the min nor max cost path for rst } n \\ \xi^{p^x n} + \xi^{p_{\max rst} n} & \text{if } p \text{ is the min cost path for rst } n \\ 0 & \text{if } p \text{ is the max cost path for rst } n \end{cases}$$

The new solution thus becomes

$$\Xi_{n+1} = (1 - \lambda)\, \Xi_n + \lambda X_n$$

where n is the iteration number, X_n is the descent direction selected for iteration n, and λ is the step length $(0 \leqslant \lambda \leqslant 1)$.

The proposed implementation of the IA algorithm is outlined in Figure 2. The figure shows that, as with the original algorithm shown in Figure 1, the practical implementation iterates between searching in the feasible assignment space, and updating of path costs according to the selected assignment; however, the practical implementation includes variations in gap functions and descent directions, which were empirically found to be effective search strategies.

The person assignment-based intermodal IA algorithm shown in Figure 2 was implemented in the VISTA framework. The implementation makes use of VISTA's Routesim traffic simulator to estimate car and bus travel times according to cell transmission model vehicle propagation relationships. Based on the simulator's spatio-temporal vehicle trajectories, bus arrival times at bus stops are then used to determine person travel times on buses, as well as bus stop waiting times. The road link travel times, bus travel times and bus stop waiting times are then used, along with fixed transfer costs, such as parking fees and bus fares, as well as distance-based pedestrian link travel times, to calculate generalized costs.

These costs are used to update path costs for all paths in the path set. Further, for path generation in Phase 1, these costs are used in an intermodal least cost path algorithm to generate intermodal paths. A new equilibrium assignment is then estimated according to the IA descent direction and line search procedure.

In this implementation, because Routesim simulates only vehicle movements, rather than person movements, it does not capture the relationships between buses and their passengers. Specifically, the effects of increased loading and dwell time are not captured, nor the effects of bus crowding and capacity limits. To illustrate, given an assignment pattern along with a set of link travel times and bus stop arrival times, the bus vehicle loading can be determined and dwell times can be assigned based on the number of boardings and alightings at each bus stop. Those dwell times are then used in the simulator

Phase 0 – Initialize
Set $n = 0$.
Set link travel time to free flow.
Calculate least cost paths for each $rstb$.
Set Ξ_0 to all-or-nothing assignment of d_{rs}^t to least cost path for $rstb$.

Phase 1 – Search in Extreme Point Direction
Simulate traffic conditions with assignment Ξ_n.
Update link travel times.
If ($n = 0$) or ((|path set|$_n$ − |path set|$_{n-1}$) > **routing-stop-percentage** ∗ |path set|$_{n-1}$)
  Calculate least cost paths for each rst.
  Add new paths to path set.
Set descent direction (all-or-nothing assignment to least cost paths, P_n).
Select step length $\lambda_n = \text{argmin}_\lambda(V_{demand}((1-\lambda)\Xi_n + \lambda P_n))$ using golden section search.
Assign demand to $\Xi_{n+1} = (1-\lambda_n)\Xi_n + \lambda_n P_n$.
If $n <$ **assignment-max-extremes**
  Repeat Phase 1.
Else
  Go to Phase 2.
Set $n = n + 1$.

Phase 2 – Search in Smith Direction
Simulate traffic conditions with assignment Ξ_n.
Update link travel times.
Choose descent direction Δ_n.
Select step length $\lambda_n = \text{argmin}_\lambda(V_{smith}(\Xi_n + \lambda\Delta_n))$ using golden section search.
Assign demand to $\Xi_{n+1} = \Xi_n + \lambda_n\Delta_n$.
If $V_{smith}(\Xi_{n+1})/V_{smith}(\Xi_n) >$ **assignment-switch-ratio**
  Repeat Phase 2
Else
  Go to Phase 3
Set $n = n + 1$

Phase 3 – Search in Non-extreme Direction
Simulate traffic conditions with assignment Ξn.
Update link travel times.
Choose descent direction X_n.
Select step length $\lambda_n = \text{argmin}_\lambda(V_{cost}((1-\lambda)\Xi_n + \lambda P_n))$ using golden section search.
Assign demand to $\Xi_{n+1} = \Xi_n + \lambda X_n$.
If ($V_{cost}(\Xi_{n+1})/\Sigma_{rstp}\psi_{rstp}(\Xi_{n+1}) >$ **cost-gap-percentage**)
  Repeat Phase 3.
  Set $n = n + 1$.
Else
  Terminate.

Note:
The routing-stop-percentage, assignment-switch-threshold, assignment-switch-ratio and cost-gap-percentage are convergence parameters that can be set within the VISTA implementation of IA.

Figure 2. The IA algorithm implemented in VISTA.

to model bus stopping behavior. Next, the link and path travel costs are updated based on the simulator results and a new assignment is calculated. However, the new assignment is based on dwell times that were associated with the previous iteration's assignment pattern. A similar problem with the offset of assignment solutions and simulation results exists when calculating crowding and capacity on buses. In other words, the inability to model person movements explicitly results in a disconnect between the interaction of passengers with bus vehicles.

In addition, the model and algorithm account for individual preferences by allowing different generalized cost functions to be defined for different traveler groups; however, no effective method of calibrating these cost functions exists. As such, the model is cannot currently be used to solve real-world intermodal assignment problems.

In general, the search procedure assumes that a path's cost increases with the number of people loaded on that path; however, as previously explained, due to path interactions, the cost function may not behave this way on actual networks. As such, the algorithm is considered heuristic and its convergence cannot be proven; however, in practice preliminary results show reasonably close approximations to equilibrium.

It should be noted that while the rate of convergence is not guaranteed, it is certain that the algorithm will not diverge. This is true because should it be found that the solution cannot be improved in a particular direction, then the step length will be set to $\lambda = 0$, such that $\Xi_{n+1} = \Xi_n$. The gap function is thus limited to $V_{n+1} \leqslant V_n$. This holds true for the gap functions of all three phases.

6. Computational Results

The person assignment-based intermodal DTA model was implemented in VISTA, and test results on a small test network of cars and buses are presented in this section. First, the test network is described along with a discussion of convergence of the IA algorithm for the test network. Next, a sample analysis of transit signal priority is presented for the test network. No tests on real-world networks are currently planned, since real-world tests will require detailed person trip data and disutility parameters, which may be difficult to obtain, calibrate and validate.

The intermodal test network, shown in Figure 3, includes 82 roadway intersection nodes and 169 roadway links. In addition, the network is traversed by five bus routes, each traveling in two directions, and bus stops are connected the nearest roadway nodes by pedestrian links to allow for intermodal transfers. Trips are generated at 26 nodes in 19 zones, and parking areas are located at all trip ends, as well as at bus stop transfer locations. In addition, 21 interior arterial and ramp intersections are signalized (see Figure 4).

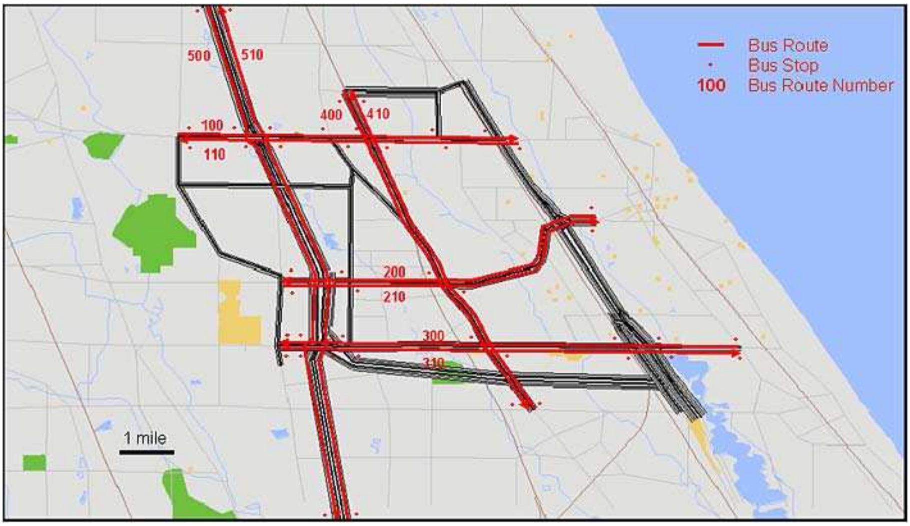


Figure 3. Intermodal Test Network.

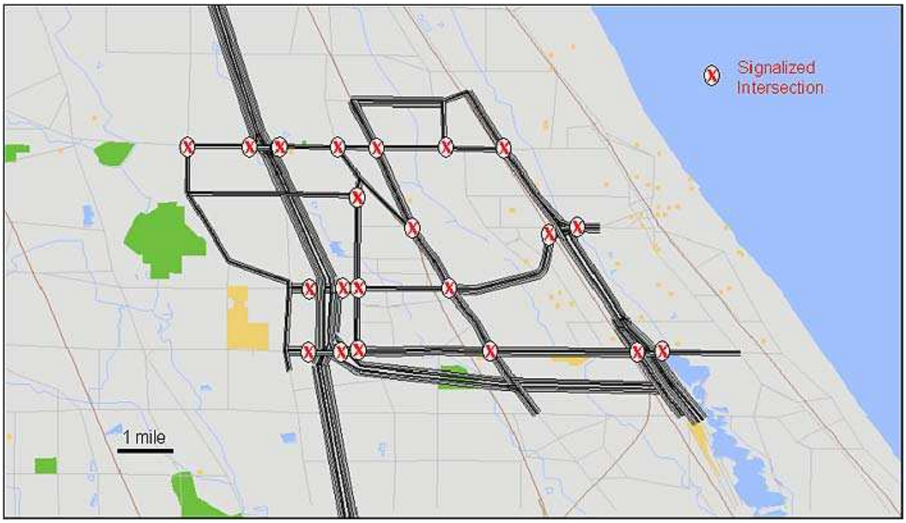


Figure 4. Signalized Intersections in the Intermodal Test Network.

A total demand of 24,341 person trips is loaded into the network in the first 1.5 hours of a 3-hour simulation period. These trips were divided among two behavior types as defined in Table 1. The costs parameters are defined such that travelers of the first behavior type prefer auto travel to bus travel. Travelers of the second behavior type consider auto travel, bus travel and bus transfer time equally onerous, but consider walking time more slightly more onerous.

Table 1. Traveler behavior types.

Behavior Type	Number of Travelers	Auto Travel Cost Factor ($ / hr)	Bus Travel Cost Factor ($ / hr)	Bus Transfer Cost Factor ($ / hr)	Pedestrian Travel Cost Factor ($ / hr)
1	12,134	5.04	5.40	5.76	7.20
2	12,207	5.40	5.40	5.40	5.76

Further, bus fares were set to $0.50 for initial boardings as well as transfers, and parking fees were $2.00 at all parking locations in the network.

The IA algorithm was performed on the test network with each phase was permitted to run until it was unable to improve its solution. Specifically, Phase 1 was permitted to generate paths until the number of new paths generated was negligible (less than 1% of the path set, in this case). Phase 1 terminated with a demand gap of 0.00, but with a cost gap of 0.75%. The demand gap value of 0.00 suggests that an equilibrium assignment has been found, but the cost gap indicates that the solution may still be improved. The apparent inconsistency

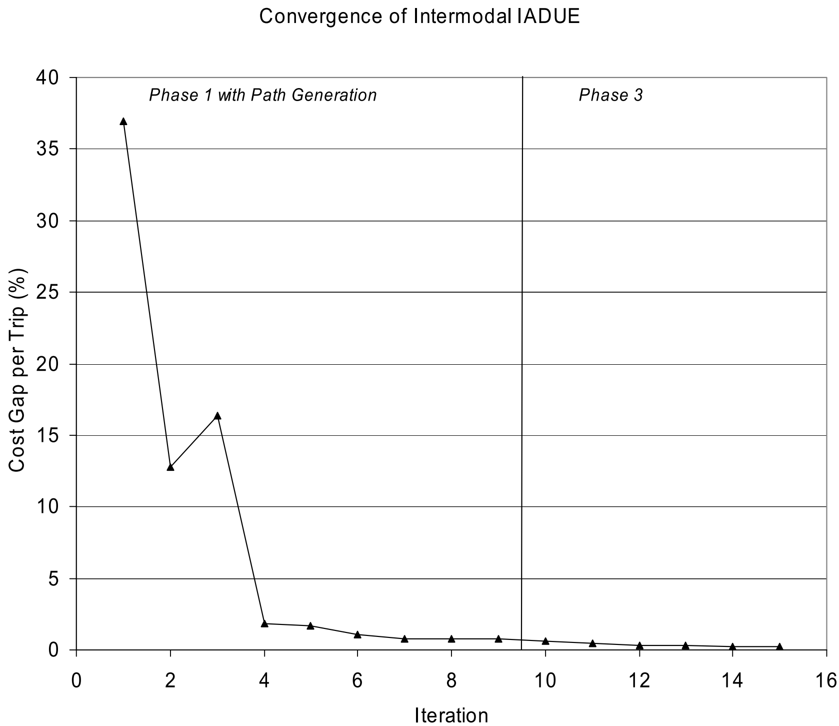


Figure 5. Convergence of the Cost Gap per Trip for the Intermodal Test.

occurs because the demand gap is calculated as a system cost, so inefficiency in one traveler's assignment may be compensated by an improvement in another traveler's assignment. In contrast, the cost gap considers travelers individually, without allowing for one traveler's inefficiency to be compensated by an improvement in another traveler's route choice. Since both Phases 1 and 2 of the IA algorithm attempt only to minimize the demand gap, the stopping criteria for both phases were satisfied by the demand gap of 0.00, the algorithm continued with Phase 3 to minimize the cost gap. Figure 5 shows the value of the average cost gap per trip with each iteration throughout the algorithm. As expected, the cost gap fluctuated throughout Phase 1. This occurred because the addition of new lower-cost paths and extreme points resulted in increases in the demand and cost gap values. The plot shows that, despite the fluctuations, the extreme direction search of Phase 1 brought the cost gap down to about 0.75%. Next, Phase 3 reduced the cost gap down to the final value of 0.25%, or approximately $0.01 per trip for an average trip cost of $2.26.

The total computational time required for the test was 15.1 hours on a machine with dual 2GHz Athlon servers with 4GB RAM; however, there were several other large processes running on the machine at the same time, so it is expected that the computational time would be much lower with a dedicated machine. Table 2 shows a breakdown of the number of iterations, the number of simulations required, and the computational time required for each phase. More specifically, in each phase, each iteration includes a line search during which several step lengths are tested. For each step length test, a simulation

Table 2. Computational Time for Test Run A.

Phase	Number of Iterations	Number of Simulations	Computational Time
1 – path generation	9	35	11:58:44
2	–	–	–
3	6	34	03:03:37
Total	15	69	15:02:21

Table 3. Line Search Parameters.

Parameter	Value
Assignment-search-low	0.38
Assignment-search-high	0.62
Assignment-stop-difference	0.05
Assignment-previous-gap-ratio	0.99
Assignment-worst-gap-ratio	0.05

run is performed. With the line search parameters set to the values shown in Table 3, each line search required 3–8 simulations, with each simulation taking an average of 28 seconds. In addition, each performance of the least cost path algorithm for path generation required an average of 2.8 minutes. The most time-consuming part of the procedure was the storing of path assignments after each iteration, a process that required over an hour each time.

In Phase 1, the routing-stop-percentage was set to 1%, such that when the number of new paths generated was less than 1% of the total number of paths in the path set, the time dependent shortest path calculation was no longer invoked for subsequent iterations. Figure 6 shows the growth of the path set through the eight path iterations of Phase 1 during which paths were generated. The plot shows that many paths were added in the early iterations, but that the growth of the path set leveled off in later iterations to a final path set size of 1091 for 425 OD pairs for 2 behavior types.

TSP Strategy 1 was provided along bus routes 100 and 110, where each route had a headway of 3 minutes. The average travel times of priority routes, listed in Table 4, show that TSP resulted in 17% and 32% improvements over base case travel times for routes 100 and 110, respectively. It would be expected

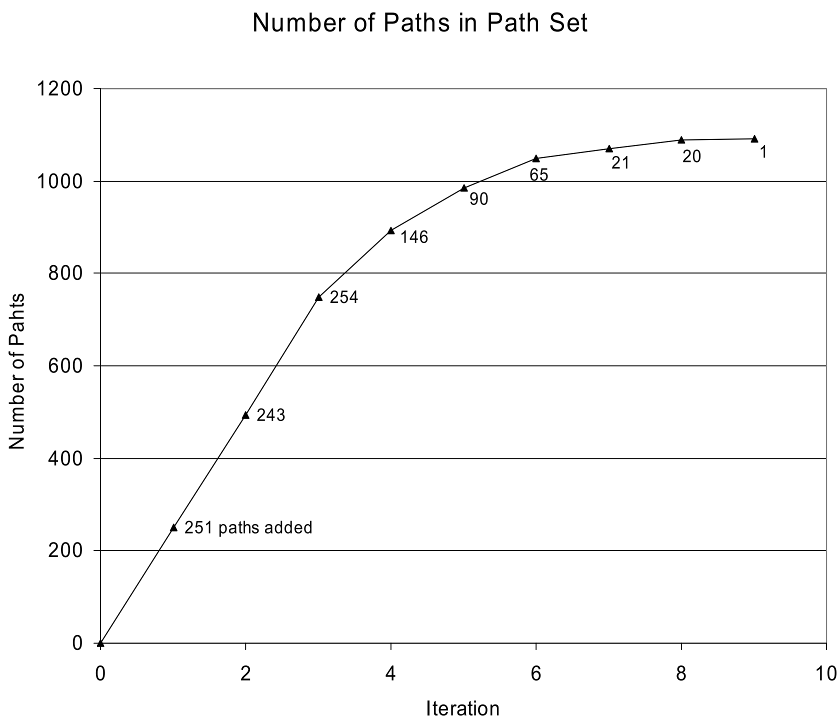


Figure 6. Growth of the Path Set.

Table 4. Bus Route Travel Times along TSP Corridor.

Route	Number of Buses	Corridor Length (miles)	Base Case Travel Time Travel Time (min)	TSP Case Travel Time Travel Time (min)	Difference (min)	Difference
100	40	3.8	7.0	5.8	-1.2	17%
110	40	3.8	9.2	6.3	-2.9	32%

Table 5. Bus Ridership on Priority Routes.

Route	Number of Buses	Base Case Bus Ridership	TSP Case Bus Ridership	Difference
100	40	32	34	2
110	40	108	115	7

Table 6. Automobile Travel Times along TSP Corridor.

Direction	Base Case Number of Autos	Test Case Number of Autos	Base Case Average Auto Speed (mph)	TSP Case Average Auto Speed (mph)	Difference (mph)	Difference
EB	2,161	2,161	31.0	34.7	3.7	12%
WB	2,476	2,476	13.9	19.4	5.6	40%

Table 7. System-wide measures.

Direction	Base Case	Test Case
Total System Cost	$50,517.50	$50,464.00
Total Person Travel Time (hrs)	1,199.2	1192.6
Total Vehicle Travel Time (hrs)	120.3	109.9
Total Vehicle Miles Traveled (mi)	163,876	160,672

that such in improvement in bus travel time would correspond with an increase in ridership for that route; however, Table 5 shows that in fact, routes 100 and 110 experienced only minimal increases in ridership.

Examination of the average automobile travel speeds on the corridor reveals that just as bus travel times improved, automobile travel speeds also increased, as shown in Table 6. An increase of 3.7mph in the EB direction and 5.6mph in WB direction resulted in little incentive to switch to bus travel.

In addition, Table 7 shows that TSP reduces system-wide costs, travel times and vehicle miles traveled. With the intermodal approach, the person travel cost is observed directly in the model, along with person travel times. In addition, vehicle-based statistics can also be observed.

7. Conclusions

This paper proposed an IA assignment algorithm to find the equilibrium path assignment for the dynamic intermodal person-trip assignment problem. The algorithm assumes continuous and monotonic cost functions. In reality, traffic signals and mode transfers may result in discontinuous travel time costs, and path interactions may cause path costs to be non-monotonic; however, despite the problems with these assumptions, preliminary tests of the algorithm show that it efficiently approaches an equilibrium solution.

The IA algorithm was implemented within the framework of the VISTA dynamic traffic assignment software. As with the original automobile assignment-based VISTA model, the intermodal person trip-based approach iterates between traffic simulation, path calculations, and network assignment; however, the intermodal path calculation finds the least cost intermodal path, where costs are defined to reflect traveler preferences regarding travel time, fares, parking costs and other factors. The interactions of cars and buses in the shared roadway network are captured by VISTA's Routesim simulator, which propagates vehicles according to cell transmission model logic. Further, turning movements, signalized intersections and transit signal priority logic are included in the simulator. The simulator also captures the added length of bus vehicles, as well as their frequent stopping behavior. Since person movements are not captured by Routesim, the interactions of buses and passengers are not captured; for example, dwell times do not reflect the number of boardings and alightings, and bus crowding and capacity limits are not modeled.

Computational results were presented to illustrate the convergence behavior and computational time required to run the intermodal IA algorithm. In addition, sample tests of TSP on the intermodal network were performed. No tests on real-world networks are currently planned, since real-world tests will require detailed person trip data and disutility parameters, which may be difficult to obtain, calibrate and validate.

In short, this approach captures interactions between cars and buses in the simulator, as well as intermodal route choices, such that bus travel times and travel time variability can be observed in the simulator output. Since mode choice is determined within the model, the impacts of transit policies on ridership are captured by the model and can be directly observed in the model output. Future work in this area includes development of effective methods of calibrating the generalized cost functions.

Acknowledgments

This research was supported by the US Department of Transportation through grant DTRS 99-G-0005 from the Midwest Regional University Transportation Center at the University of Wisconsin-Madison.

References

Aashtiani, Hedayat Zokaie. 1979. *The Multimodal Traffic Assignment Problem.* Ph.D. dissertation, Massachusetts Institute of Technology.

Abdelghany, K.F., A.F. Abdelghany, H.S. Mahmassani, and A.S. Abdelfatah. 2001. Evaluating Bus Preemption Strategies at Signalized Intersections Using a Multimodal Dynamic Network Assignment-Simulation Methodology. *80th Annual Meeting of the Transportation Research Board*, 2001. Pre-print CD ROM.

Abdelghany, K.F. and H.S. Mahmassani. 2001. Dynamic trip assignment-simulation model for intermodal transportation networks. *Transportation Research Record* 1771: 52–60.

Abdulaal, M. and L. LeBlanc. 1979. Methods for combining modal split and equilibrium assignment models. *Transportation Science* 13 (4): 292–314.

Abrahamsson, T. and L. Lundqvist. 1999. Formulation and estimation of combined network equilibrium models with applications to Stockholm. *Transportation Science* 33 (1): 80–100.

Benjamins, M., D.R. Lindveld and R. van Nes. 2002. Multimodal travel choice modelling: A supernetwork approach. *81st Annual Meeting of the Transportation Research Board*, 2002. Pre-print CD ROM.

Boile, M.P., L.N. Spasovic, and A.K. Bladikas. 1995. Modeling Intermodal Auto-Rail Commuter Networks. *Transportation Research Record* 1516: 38–47.

Carlier, K., S. Fiorenzo-Catalano, C. Lindveld and P. Bovy. A supernetwork approach towards multimodal travel modeling. *82nd Annual Meeting of the Transportation Research Board*, 2003. Pre-print CD ROM.

Dafermos, S.C. 1980. Traffic equilibrium and variational inequalities. *Transportation Science* 14 (1): 42–54.

Dafermos, S.C. 1982. The general multimodal network equilibrium problem with elastic demand. *Networks* 12: 57–72.

Daganzo, C.F. 1994. The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research* 28B (4): 269–287.

de Cea, J., J.E. Fernandez, V. Dekock. 2003. ESTRAUS: A computer package for solving supply-demand equilibrium on multimodal urban transportation networks with multiple user classes. *82nd Annual Meeting of the Transportation Research Board*, 2003. Pre-print CD ROM.

Fernandez, E., J. de Cea, M. Florian and E. Cabrera. 1994. Network equilibrium models with combined models. *Transportation Science* 28 (3): 182–92.

Florian, M. 1977. A traffic equilibrium model of travel by car and public transit modes. *Transportation Science* 11 (2): 166–79.

Florian, M. and S. Nguyen. 1978. A combined distribution modal split and trip assignment model. *Transportation Research* 12: 241–46.

Florian, M. and H. Spiess. 1983. On binary mode choice/assignment models. *Transportation Science* 17 (1): 32–47.

Los Alamos National Laboratory. 2002. *TRANSIMS: Transportation Analysis Simulation System, Version: TRANSIMS – 3.0*. Vol. 3, *Modules*. Report no. LA-UR-00-1725. Los Alamos.

Mahmassani, H.S., S. Peeta, T.Y. Hu and A.K. Ziliaskopoulos. 1993. Dynamic traffic assignment with multiple user classes for real-time ATIS/ATMS applications. In *Proceedings of the Advanced Traffic Management Conference, St. Petersburg, Florida*. Washington D.C.: Federal Highway Administration.

Mahmassani, H.S. 2001. Dynamic network traffic assignment and simulation methodology for advanced system management applications. *Networks and Spatial Economics* 1: 267–92.

Smith, M.J. 1979. The existence, uniqueness and stability of traffic equilibria. *Transportation Research* 13B (4): 295–304.

Smith, M.J. 1983. The existence and calculation of traffic equilibria. *Transportation Research* 17B (4): 291–303.

Wu, Z.X. and William H.K. Lam. 2003. A combined modal split and stochastic assignment model for congested networks with motorized and non-motorized transport modes. *82nd Annual Meeting of the Transportation Research Board*, 2003. Pre-print CD ROM.

Ziliaskopoulos, A.K. and S.T. Waller. 2000. An internet based geographic information system that integrates data, models and users for transportation applications. *Transportation Research* 8C (1): 427–44.

Chapter 11

Modeling the Transient Nature of Dynamic Pricing with Demand Learning in a Competitive Environment

Soulaymane Kachani,* Georgia Perakis,† and Carine Simon‡

**IEOR Department, Columbia University, New York, NY,*
E-mail: kachani@ieor.columbia.edu

†MIT Sloan School of Management, Cambridge, MA, E-mail: georgiap@mit.edu

‡MIT Operations Research Center, Cambridge, MA, E-mail: casimon@mit.edu

Abstract This paper focuses on joint dynamic pricing and demand learning in an oligopolistic market. Each firm seeks to learn the price-demand relationship for itself and its competitors, and to set optimal prices, taking into account its competitors' likely moves. We follow a closed-loop approach to capture the transient aspect of the problem, that is, pricing decisions are updated dynamically over time, using the data acquired thus far.

We formulate the problem faced at each time period by each firm as a Mathematical Program with Equilibrium Constraints (MPEC). We utilize variational inequalities to capture the game-theoretic aspect of the problem. We present computational results that provide insights on the model and illustrate the pricing policies this model gives rise to.

Keywords: dynamic pricing; demand learning; variational inequalities; game theory

1. Introduction

1.1 Brief Presentation of the Problem, Motivation and Application Areas

Determining the right price to charge customers requires a company to have a wealth of information and data. In particular, the company needs information concerning the customer base, the company's own cost structure, as well as information concerning the competition and the market itself. Furthermore, it requires that prices can be adjusted in a timely fashion at minimal cost. Until recently, neither was possible. As a result, traditional pricing techniques were often static.

The fast development of information technology and e-commerce had a dramatic impact on the market place. Thanks to these tools, the sellers can gather information about customers and competitors; they can also update prices dynamically at low cost and hence, they allow the sellers to implement dynamic price optimization.

Early applications of dynamic pricing include industries where short term supply is hard to change, such as airlines, cruise boats, hotels, electricity markets. Other industries then realized the benefits of dynamic pricing strategies including retailers in brick-and-mortar stores as well as online, so much so that many companies now resort to dynamic price optimization solution providers, such as DemandTec, Khimetrics, Manugistics, Oracle, Pros Revenue Management and Sabre.

1.2 Literature Review

In recent years, revenue management and particularly pricing have drawn increased interest from both practitioners and researchers in many fields.

Revenue management is an extensively studied field, for which thorough reviews are available. Hence we refer the reader to these reviews, and focus more on the pricing and learning literature.

1.2.1 Revenue Management and Pricing Literature

The recent book by Talluri and van Ryzin (2004) provides a thorough review of the theory and practice of Revenue Management. Review papers include McGill and van Ryzin (1999), Bitran and Caldentey (2002). They provide an overview of pricing models in Revenue Management, whereas the survey paper of Weatherford and Bodily (1992) concentrates on Revenue Management in the airline industry.

Elmaghraby and Keskinocak (2003) provide a research overview of dynamic pricing. They observe that three main characteristics of the market environment influence the pricing problem: first, whether replenishment of inventory is allowed; second, whether demand arrivals are independent over time; third, whether customers act myopically or strategically.

Dynamic Pricing models with no replenishment, and independent demands over time rely on common assumptions: a market with imperfect competition (e.g monopoly), a finite selling horizon with finite stock and no replenishment. The demand is typically decreasing in price. The goal of the firm in these settings is to maximize the expected profits over the selling horizon.

Gallego and van Ryzin (1997) and Feng and Gallego (1995) model the demand as a Poisson process with a rate that is decreasing in price. Bitran et al. (1998) as well as Bitran and Mondschein (1997) consider a demand rate which depends on time and has a known distribution. Lazear (1986), Elmaghraby et al. (2002) model the demand using reference prices; Lazear's demand model is deterministic whereas Elmaghraby et al. (2002) assume that demand is stochastic with a known distribution. Achabal and Smith (1998) model demand as depending on price, time and inventory level. Maglaras and Meissner (2004) examine two problems: dynamic pricing in a monopoly with imperfect competition and dynamic capacity control with exogenous prices. They show that these problems have a common formulation as a single resource, single product pricing problem. In all these papers except for Elmaghraby et al. (2002), the customers are assumed to act myopically. Lazear (1986) and Elmaghraby et al. (2002) study periodic pricing policies where prices are updated at discrete time intervals, whereas Gallego and van Ryzin (1997) and Bitran and Mondschein (1997) study continuous time pricing policies. Some models restrict themselves to a fixed number of price changes, as in Feng and Gallego (1995), Bitran and Mondschein (1997) and Feng and Xiao (2000).

For models of dynamic pricing with inventory replenishment, independent demand and myopic customers, most of the research focuses on a monopoly market, in a single or multi-product setting. Whereas the typical Inventory Management research considers price to be static, and exogenous, the following papers consider joint inventory management and pricing. Federgruen and Heching (1999), Zabel (1970), and Thowsen (1975) address the optimal inventory and pricing of a seller who faces uncertain demand and changes its prices periodically. They find that a base stock list price policy is optimal in a wide range of settings. Rajan et al. (1992) focus on changes that occur within an order cycle for a firm selling perishable products. Popescu and Wu (2005) study dynamic pricing for customers with repeated interactions: in this setting, customers are sensitive to the pricing history through a reference price. They show that the pricing strategy has long term implications, in that promotions which increase short-term profits, may decrease future profits. Adida and Perakis (2005) propose a nonlinear fluid model for joint dynamic pricing and inventory control with no backorders.

Competition was studied extensively in the traditional Economic literature. The book by Friedman (1983) presents the theory of oligopoly, and Vives (1999) provides a modern theory of oligopoly using the new tools of Game Theory. Fudenberg and Tirole (1986), as well as Maskin and Tirole (1988) study dynamic oligopoly. Sweezy (1939) conjectures a kinked-demand curve in competitive oligopoly. Stigler (1947) also focusses on the kinky demand curve and shows the stickiness of prices in an oligopoly. Recent work in dynamic pricing considers competitive settings: Dockner and Jørgensen (1988) consider optimal pricing strategies in an oligopoly market, but from a marketing perspective. Bernstein and Federgruen (1999) built an inventory model for supply chains in an oligopoly, where the decision variables include prices, service levels and inventory control. Kachani and Perakis (2002) propose a deterministic fluid model for dynamic pricing in a capacitated, make-to-stock manufacturing system. Perakis and Sood (2005) propose a dynamic pricing model and study Nash equilibria in an oligopoly for perishable products.

1.2.2 Dynamic Pricing with Learning

When demand is uncertain, it is natural that the seller tries to learn it over time. Hence dynamic pricing approaches often address the issue of learning the demand function. Carvalho and Puterman (2004), and Carvalho and Puterman (2004) consider the problem of dynamic pricing when only the form of the demand function is known, not the parameters. They assume a prior distribution on the parameters, and update the parameters over time by using Kalman filters. Aviv and Pazgal (2004) propose a Markov-modulated demand model. In such a model, the state of the market encompasses all relevant information concerning the demand. In a partially-observed Markov Decision Process, the information about the state is incomplete. The seller starts with some prior information concerning the parameters of the state, and update their beliefs through a Bayesian update scheme. In Aviv and Pazgal (2004), Aviv and Pazgal consider a continuous-time model, where demand has a prior distribution. They show that there is a trade-off between a low price which yields a loss in revenue, and a high price which lowers the probability of purchase, and slows learning. Dada and Petruzzi (2002) consider a finite horizon, discrete time problem for perishable product, with restocking. At each time period, both supply and demand are decision variables. Demand is a deterministic function of price, with initially unknown parameters which are assumed to follow a prior distribution. Boyd and Lobo (2003), justify price variations in the market by the rational learning behavior of the firms. They consider a monopoly with stochastic linear demand whose parameters are not known but assume a prior distribution. Balvers and Cosimano (1990) focus on a monopoly with stochastic linear demand with unknown intercept and slope. The slope is assumed to have a persistent effect, and thus prompts learning. They define the speed of learning, which is controlled by the firm since it depends on the price and show that learning implies muted responses to changes in demand or market price. Rustichini and Wolinsky (1995) study a monopoly which faces an uncertain demand, and learns about it through its pricing experience. The demand curve facing the monopoly is not constant and differs from the informed monopoly's policy. They show that even when the rate at which the demand varies is negligible, the stationary probability that the monopoly's policy deviates from the full information counterpart is non negligible. Mirman et al. (1995) examine a monopoly in a two-period horizon. They develop conditions under which the firm will find it optimal to adjust its initial price or quantity away from their myopic level in order to increase informativeness of observed market outcomes and thus increase future expected profits. Bertsimas and Perakis (2006) address dynamic pricing in a monopoly and a duopoly, where demand is a priori unknown, but learned over time.

Finally, learning is also addressed in the statistics and decision theory literatures: Kalyanam (1996) proposes a model that draws on Bayesian estimation, inference and decision theory to learn uncertain demand. Barto and Sutton (1998) provide an introduction to reinforcement learning, with wide application areas. Learning also arises in stochastic processes when the parameters are unknown, as in Easley and Kiefer (1988).

1.3 Contributions

This work builds up on the work by Bertsimas and Perakis (2006). The main contributions of this work are the following:

- Unlike most of the models in the literature concerning dynamic pricing and learning which deals with a monopoly, the model we introduce explicitly incorporates competition and uses ideas from Game Theory to compute market equilibria.
- Furthermore, our approach addresses the capacitated case: each firm has a limited inventory without replenishment.
- The bulk of the literature on dynamic pricing and learning either assumes that the demand follows a certain probability distribution, or that the seller has a certain prior distribution on the demand. These assumptions seem too strong, and therefore, our model constitutes an attempt to relax them.
- We follow a closed-loop approach, since pricing decisions are updated periodically to take into account new information as it becomes available at each time period. On the other hand, open loop policies set prices for each time period once and for all at the beginning of the selling horizon, without periodic review.
- The goal of this research is to design an approach which enables the seller to learn the price-demand relationship. As a result, this approach should yield better estimates than an open-loop policy.
- From a mathematical perspective, we formulate the problem faced at each time period by each firm as a Mathematical Program with Equilibrium Constraints (MPEC).
- We provide some computational results, and build some intuition in the 2 firm-2 period case.

1.4 Structure of the Paper

The remainder of the paper is organized as follows: In Section 2, we present the Dynamic Pricing Problem faced by each seller at each time period. We introduce the main characteristics of the problem and the demand model. In Section 3, we show that the dynamic pricing problem can be decomposed into three steps; first, assuming the price-demand parameters known, each seller would like to find the Nash equilibrium demands that would emerge on the market. Second, using the optimal demands computed in Step 1, each seller seeks to estimate the price-demand parameters for himself and his competitors. Finally, once the optimal demands and parameters are computed, each firm can set optimal prices for future periods, as well as find its competitors' optimal pricing policy. In Section 4, we show that the transient dynamic pricing problem can be formulated as a Mathematical Problem with Equilibrium Constraints (MPEC). In Section 5, we focus on the case of a duopoly with 2 time periods, and derive closed form solutions to the problem. In Section 6, we try to get some insights into the closed form solutions and present some computational results of the joint dynamic pricing and learning approach. Finally, we conclude with a discussion of the contributions of the approach.

2. Presentation of The Dynamic Pricing Problem

2.1 Main Characteristics of the Problem

The following features characterize the pricing problem we consider in this paper:

- The prices are set periodically over a finite selling horizon $\{1, \ldots, T\}$. That is, at the beginning of each period $t \in \{1, \ldots, T\}$, each firm sets its prices for the period p_i^t.
- We consider a single, perishable product with finite inventory without replenishment. That is, at the end of the selling horizon T, all unused inventory is lost. Furthermore, we restrict our study to products such that the marginal cost of an extra unit of demand is sufficiently small. This allows us to focus on revenue maximization rather than profit maximization. Yet, this assumption is not critical.
- The market we consider is a noncooperative oligopoly characterized by a few firms on the supply side, and a large amount of buyers on the demand side. Therefore, the profit of each firm depends on the prices set by all the competing firms.
- Competition in the market can be described by a Cournot model, where prices are determined by the allocations, or demands of the firms. This is indeed the model commonly used to describe markets where short-term supply is hard to adjust, and therefore holds for such perishable products as airline tickets, hotel bookings, electricity markets, etc.

- The price-demand relationship is unknown a priori, but learned over time. Only the functional form of the demand is assumed to be known. We assume parametric families of demand functions characterized by coefficients called price elasticities. For tractability purposes, we focus in this paper on linear price-demand relationships. Nevertheless, this work can be extended to nonlinear families of price-demand functions.
- We assume that we can characterize the behavior of each firm in the market (e.g price follower or optimizer). In the remainder of the paper, we will consider all firms to be revenue maximizers; this implies that we focus on revenue and not profit maximization. In other words, their pricing policy is to maximize their total revenue over the entire selling horizon.

2.2 Notations of the Model

We introduce the following notations which will be useful in the remainder of the paper:

- T: selling horizon after which all unused capacity is lost;
- $C_i, i = 1, \ldots, N$: finite inventory of each firm over the entire selling horizon; we call this inventory the total capacity of the firm.
- For a vector x with components x_i^t, $i = 1, \ldots, N$, $t = 1, \ldots, T$, we denote $x_i = (x_i^1, \ldots, x_i^T)$ the subvector corresponding to firm i, and $x^t = (x_1^t, \ldots, x_N^t)$ the subvector corresponding to time period t.
- $\widehat{p_i^t}$: price set by firm i at period t; $p_i^{t,h}$: historical price set by firm i at period t;
- d_i^t: market share, of firm i at period t; β_i^t: vector of price sensitivities of firm i at period t: it has components $\beta_{i0}^t, \beta_{i1}^t, \ldots, \beta_{iN}^t$: β_{i0}^t denotes the intercept of the demand function, i.e the demand faced by firm i when all firms set their price to 0, and can be interpreted as the total demand for firm i; β_{ii}^t is the sensitivity of firm i's demand to its own price; β_{ij}^t, $j \neq i$ are the price sensitivities of firm i's demand to its competitors' prices.
- $[\beta]$: matrix of price sensitivities for all the firms, and in all time periods; its coefficients are β_{ij}^τ for $i, j = 1, \ldots, N$, $\tau = 1, \ldots, T$.
- $d_i^t(p_i^\tau, p_{-i}^\tau)$: demand function of firm i at period t. Its arguments are p_i^t and p_{-i}^t, where the index $-i$ denotes all the competitors of firm i, i.e $\{j = 1, \ldots, N \mid j \neq i\}$.

2.3 The Demand Model

One assumption of our model is that the demands as a function of the prices belong to a parametric family. The demands are assumed to be independent over time, that is, the demand in period t solely depends on the prices set in that period. In the rest of the paper, for the sake of simplicity, the true demands are assumed to be linear functions of the prices:

$$d_i^t(p_i^t, p_{-i}^t) = \beta_{i0}^t - \beta_{ii}^t p_i^t + \sum_{j=1, j \neq i}^{N} \beta_{ij}^t p_j^t$$

The price-demand relationship for all firms and in all periods can therefore be summarized by the matrix formula:

$$d = [\beta]p + \vec{\beta_0}$$

Moreover, we assume that firm i's demand at t is a decreasing function of firm i's current price p_i^t and an increasing function of its competitors's current price p_{-i}^t. In other words, the price sensitivities β_{ii}^t and β_{ij}^t, $\forall j \neq i$ are positive. This means that the demand faced by firm i decreases as i increases its prices, and increases when its competitors increase theirs.

In the remainder of the paper, we will work with demands as our decision variables. This is indeed the standard approach in markets for which short-term supply is hard to adjust, such as airlines, electricity markets, etc. It therefore applies to our setting of perishable product, with no replenishement of inventory. Furthermore, this allows us to best exploit the information that is available to each firm as well as the particular structure displayed by the problem when written in demand variables.

We therefore need to state conditions under which the price-demand relationship $d = [\beta]p + \vec{\beta_0}$ can be inverted. This is equivalent to finding conditions sufficient for the matrix of estimates of the price sensitivities $[\beta]$ to be non-singular. In that case, we denote by $[\alpha]$ its inverse. We refer to the parameters α_{ij}^t as the price-demand parameters since they characterize the price-demand relationship. More specifically, the parameters satisfy the following constraints:

1. $\alpha_{ii}^t > 0, \ \forall\, i = 1, \ldots, N \ \forall t = 1, \ldots, T$
2. $|\alpha_{ij}^t| > 0, \ \forall\, i, j = 1, \ldots, N, \ i \neq j \ \forall t = 1, \ldots, T$
3. $\alpha_{ii}^t > \sum_{j \neq i} |\alpha_{ij}^t|, \ \forall\, i = 1, \ldots, N, \ \forall t = 1, \ldots, T$
4. $\alpha_{ii}^t > \sum_{j \neq i} |\alpha_{ji}^t|, \ \forall\, i = 1, \ldots, N, \ \forall t = 1, \ldots, T$

PROPOSITION 2.1. *If the parameters α satisfy constraints 1 to 4, then the matrix $[\alpha]$ is invertible.*

PROOF. Assumptions 1 to 4 concerning the price sensitivities imply that $\forall i = 1, \ldots, N, \ \forall \tau = 1, \ldots, T$, we have:

$$\alpha_{ii}^\tau > \sum_{j \neq i} |\alpha_{ij}^\tau|$$

$$\alpha_{ii}^{\tau} > \sum_{j \neq i} |\alpha_{ji}^{\tau}|$$

Therefore, the matrix $[\alpha]$ is strictly diagonally dominant, hence positive definite. Consequently, it is invertible. □

Therefore, the price-demand relationship can be written as:

$$p = \vec{\alpha_0} - [\alpha]d + \epsilon$$

or equivalently:

$$p_i^t(d_i^t, d_{-i}^t) = \alpha_{i0}^t - \alpha_{ii}^t d_i^t - \sum_{j=1, j \neq i}^{N} \alpha_{ij}^t d_j^t$$

where ϵ_i^t is a random noise. The firm makes the following distributional assumption on the random noise: ϵ_i^t are independent random variables with a gaussian distribution $N(0, (\sigma_i^t)^2)$. An unbiased estimate of the variance of each parameter is (see Rice, 1995):

$$s^2(\alpha_{ij}^{t-1}) = \frac{\sum_{\tau=1}^{t} (p_i^{\tau} - \alpha_{i0}^{\tau} + \sum_{j=1}^{N} \alpha_{ij}^{\tau} d_j^{\tau})^2}{t - N - 1}$$

since $N + 1$ parameters (α_{ij}^{τ} $j = 0, \ldots, N$) have already been estimated from the data.

In addition, in order to allow more flexibility in the modeling of the demand, we consider parameters which slowly vary in time:

$$\|\alpha_i^t - \alpha_i^{t+1}\| \leqslant \delta_i \|\alpha_i^t\|, \ \forall\, i = 1, \ldots, N, \ \forall\, \tau = 1, \ldots, T-1$$

where δ_i are prespecified constants called volatility parameters, which impose the condition that the parameters α_{ij}^t are Lipschitz continuous.

For instance, the usual regression condition sets $\delta_i = 0$, which implies that the parameters are constant in time.

Finally, the parameter estimates must be chosen such that the prices they define are nonnegative:

$$\alpha_{i0}^{\tau} - \sum_{j=1}^{N} \alpha_{ij}^{\tau} d_j^{\tau} \geqslant 0 \quad \forall\, i = 1, \ldots, N, \ \forall\, \tau = 1, \ldots, T$$

$$\forall d \geqslant 0 \text{ s.t} \qquad \sum_{\tau=1}^{T} d_i^{\tau} \leqslant C_i \qquad \forall\, i = 1, \ldots, N$$

In order to ensure the strict inequalities in constraints 1 to 4, we make the following assumptions:

ASSUMPTIONS 1. *There exists* $\varepsilon > 0$ *such that:*

1. $\alpha_{ii}^t \geqslant \varepsilon, \ \forall\, i = 1, \ldots, N, \ \forall t = 1, \ldots, T$
2. $|\alpha_{ij}^t| \geqslant \varepsilon, \ \forall\, i, j = 1, \ldots, N, \ \forall t = 1, \ldots, T$
3. $(1-\varepsilon)\alpha_{ii}^t \geqslant \sum_{j \neq i} |\alpha_{ij}^t|, \ \forall\, i = 1, \ldots, N, \ \forall t = 1, \ldots, T$
4. $(1-\varepsilon)\alpha_{ii}^t \geqslant \sum_{j \neq i} |\alpha_{ji}^t|, \ \forall\, i = 1, \ldots, N, \ \forall t = 1, \ldots, T$

2.4 The Information Structure

In this subsection, we specify what information is available to each firm at the beginning of each period.

- We assume that at the beginning of the selling horizon, the firms know the total capacities of each of the sellers in the market $C_i, \ \forall i = 1, \ldots, N$. For instance, an airline who competes on a given flight knows the number of seats and the number and type of plane its competitors allocated to that flight; in the hotel industry, hotel managers know the total number of rooms of their competitors; car rental companies may have access to the number and type of cars used by their competitors, etc.
- Furthermore, at each time period t, each firm observes the prices that were set in the previous period $p_1^{t-1}, \ldots, p_N^{t-1}$. This is a realistic assumption in a lot of markets, such as airlines tickets, leisure goods such as hotel, cruise, equipment rentals, etc. Indeed, nowadays, information technology and the internet enable both companies and customers to have access to a wealth of prices. For instance, websites such as Expedia, Travelocity, Orbitz, etc display the prices of all airlines competing for a given flight; similar websites exist for concert, hotel, cruise tickets, as well as equipment rentals.
- Each firm also observes the demand that it captured in the previous period, i.e. d_i^{t-1}. However, in order to be more realistic, we do not assume that a firm observes its competitors' realized demands. For instance, hotel managers do not have access to the vacancies of their competitors, airlines do not know if their competitors' flights departed with empty seats, etc.

2.5 Objectives of Each Firm

At the beginning of each time period t, given the data available, that was specified in the previous section, each firm seeks to:

- Estimate the price-demand parameters of all the firms: $\alpha_i, \ \forall i = 1, \ldots, N$.
- Find the market-equilibrium demands corresponding to these parameter estimates.

- Finally, each firm wants to use its estimates of the parameters and the demands to set its own future prices so as to maximize its total revenue over the entire selling horizon, while guessing its competitors' optimal pricing policy.

3. The Transient Dynamic Pricing Problem

In this section, we focus on the transient problem, which each firm faces at the beginning of selling period $t \in \{2, \dots, T\}$. As outlined in the previous section, the firm seeks to estimate the price-demand parameters, in order to be able to compute the market-equilibrium demands, and set its prices for future periods, while guessing its competitors' optimal policy. The problem can therefore be decomposed into three steps. In Step 1, we will consider the problem of finding the market-equilibrium demands, assuming the price-demand parameters are known. In Step 2, we address the problem of estimating the price-demand parameters for fixed demands. Finally, in Step 3, we turn to the price setting for future periods.

3.1 Step 1: Computation of the Demands

First, in what follows, we will take the view point of an external observer of the market. In other words, each firm is on an equal footing, and has the same data available about itself and its competitors. Thus, at this point, we do not incorporate in our model the fact that each firm observes its own past demand. This enables us to take advantage of the symmetry of information between the firms.

Furthermore, we assume that the price parameters α are known. Each firm i seeks to find the Nash equilibrium demands that emerge in the market. As each firm is a revenue maximizer, then for fixed demands of its competitors d_{-i}, its best response demands are those which maximize its total revenue over the entire selling horizon. Each firm therefore solves a best response problem $\mathcal{BR}_i(d_{-i}, \alpha_i)$:

$$\max_{d_i} \Pi_i(d_i) = \sum_{\tau=1}^{T} d_i^\tau \left(\alpha_{i0}^\tau - \sum_{j=1}^{N} \alpha_{ij}^\tau d_j^\tau \right) \tag{1}$$

$$\text{s.t} \qquad \sum_{\tau=1}^{T} d_i^\tau \leqslant C_i \tag{2}$$

$$\alpha_i^\tau - \sum_{j=1}^{N} \alpha_{ij}^\tau d_j^\tau \geqslant 0, \qquad \forall \tau = 1, \dots, T \tag{3}$$

$$d_i \geqslant 0 \tag{4}$$

The objective function $\Pi_i(d_i) = \sum_{\tau=1}^{N} d_i^\tau p_i^\tau(d_i^\tau, d_{-i}^\tau)$ is the total revenue of firm i over the entire time horizon, and in (1), the price-demand relationship has been replaced by its actual expression. Constraint (2) incorporates the fact that the problem we are addressing is capacitated, that is, the total demand captured by the firm cannot exceed its overall capacity. Constraint (3) states that the prices defined by the price-demand relationship should be nonnegative. Finally, constraint (4) is the nonnegativity of the demands.

Notice that in this subsection, we compute the demands assuming that the price-demand parameters α are known. Subsection 3.2 deals with the estimation of these parameters.

Furthermore, each firm solves a similar problem for all its competitors, in order to determine their best response policies. This approach is reasonable since we have assumed that all the firms are revenue maximizers. Hence, in order to compute the market equilibrium demands, each firm simultaneously solves N best response problems: $\mathcal{BR}_i(d_{-i}, \alpha_i),\ \forall\, i = 1, \ldots, N$.

The solution to these N simultaneous optimization problems determines the market equilibrium demands for each firm, at each time period, as functions of the price-demand parameters. These demands are the Nash equilibrium demands in the market.

We call $\mathcal{BR}(\alpha)$ the optimization problem corresponding to the computation of the Nash equilibrium demands as a function of the parameters α.

In Section 4.1, we establish existence of the Nash equilibrium under the conditions 1 to 4 on the parameters stated above.

3.2 Step 2: Estimation of the Parameters

If the firms somehow knew the demands on the market, they could estimate their own price-demand parameters, as well as their competitors': each firm i seeks the price estimate given by the price function: $p_i^\tau(d_i^\tau, d_{-i}^\tau) = \alpha_{i0}^\tau - \sum_{j=1}^{N} \alpha_{ij}^\tau d_j^\tau$ to be the best estimate of the observed price $\hat{p}_i^\tau$ for all past periods $\tau = 1, \ldots, t-1$. Thus it minimizes the error between the above price function, and the past market prices that were observed. To perform this estimation, one may use either the absolute value error:

$$\left| \widehat{p_i^\tau} - \alpha_{i0}^\tau + \sum_{j=1}^{N} \alpha_{ij}^\tau d_j^\tau \right|$$

or the squared error:

$$\frac{1}{2}\left(\widehat{p_i^\tau} - \alpha_{i0}^\tau + \sum_{j=1}^{N} \alpha_{ij}^\tau d_j^\tau \right)^2$$

Note that for known demands, the problem of estimating the price-demand parameters by minimizing the squared error is the traditional linear regression or least-squares estimation.

However, in reality, the demands are not known, but the Nash equilibrium demands were computed in Step 1 as a function of the parameters α. Due to the capacity constraints, which make each best-response problem nonseparable in time, the Nash equilibrium demands are functions of the parameters for all time periods 1 to T. Therefore, the error term $\frac{1}{2}(\widehat{p_i^\tau} - \alpha_{i0}^\tau + \sum_{j=1}^N \alpha_{ij}^\tau d_j^\tau(\alpha))^2$ for $\tau < t$ involves parameters α^1 through α^T. As a result, the firms also need to estimate future price parameters α^t to α^T, as a function of past parameters α^1 through α^{t-1}.

If the parameters are stationary, then a good estimate of future parameters is an average of all previous parameters:

$$\forall \tau = t+1, \ldots, T \quad \alpha_i^\tau = \sum_{\theta=1}^{\tau} \omega_i^\theta \alpha_i^\theta, \quad \text{where } \sum_{\theta=1}^{\tau} \omega^\theta = 1$$

If, on the other hand, the parameters are expected to increase or decrease in time, then an ARIMA(0,1,1) process might capture the trend effect. In the rest of the paper, for the sake of simplicity, we will work with time average of past parameters. Therefore, each firm needs to solve the following N optimization problems: $\forall\, i = 1, \ldots, N$:

$$\min_{\alpha_i} \quad \frac{1}{2} \sum_{\tau=1}^{t-1} \left(\widehat{p_i^\tau} - \alpha_{i0}^\tau + \sum_{j=1}^N \alpha_{ij}^\tau d_j^\tau \right)^2 \tag{5}$$

$$\text{s.t} \quad \alpha_{ii}^\tau \geqslant \varepsilon \qquad \forall\, i = 1, \ldots, N,\ \tau = 1, \ldots, T \tag{6}$$

$$(1-\varepsilon)\alpha_{ii}^\tau \geqslant \sum_{j \neq i} |\alpha_{ij}^\tau| \qquad \forall \tau = 1, \ldots, T \tag{7}$$

$$(1-\varepsilon)\alpha_{ii}^\tau \geqslant \sum_{j \neq i} |\alpha_{ji}^\tau| \qquad \forall \tau = 1, \ldots, T \tag{8}$$

$$\|\alpha_i^\tau - \alpha_i^{\tau+1}\| \leqslant \delta_i \|\alpha_i^\tau\| \qquad \forall \tau = 1, \ldots, T-1 \tag{9}$$

$$\alpha_i^\tau = \frac{1}{\tau} \sum_{\theta=1}^{\tau} \alpha_i^\theta \qquad \forall \tau = t+1, \ldots, T \tag{10}$$

$$d_j = (d_j^1, \ldots, d_j^T) \in \mathcal{BR}_j(d_{-j}, \alpha_j) \quad \forall j = 1, \ldots, N \tag{11}$$

Note that the objective function (5) can be equivalently written as the absolute value error. Constraints (6), (7), (8) and (9) correspond to conditions 1 to 4 that were imposed on the parameters to guarantee that the price-demand relationship is invertible. Constraint (10) reflects the fact that the parameters are allowed to vary slowly in time. Constraint (11) is the estimate of future parameters as an average of the past parameters. Finally, constraint (12) captures the fact that the demands are the market equilibrium demands computed in Step 1.

Furthermore, for known demands, the estimation problem of firm i solely involves firm i's parameters. Each estimation problem thus yields the set of optimal parameters for that firm, as a function of the demands of all firms.

However, the Nash equilibrium demands computed in Step 1 are functions of all firms' parameters. Hence when solving for the Nash equilibrium demands, each estimation problem is therefore coupled to the others. As a consequence, the N estimation problems need to be solved simultaneously. The simultaneous solution to the N problems yields the set of optimal price-demand parameters for all firms and for all time periods.

3.2.1 An Alternative Estimation Problem

Let us assume that on top of observing the prices on the market, we have price data available for $\tau = 1, \ldots, T$, for instance, from a database of historical prices. Let $p_i^{\tau,h}$ be an estimate of p_i^τ based on historical data. Then at the beginning of period t, each firm may use the prices $\widehat{p_i^\tau}$ it observed for $\tau = 1, \ldots, t-1$ in order to estimate past parameters, along with its historical prices $p_i^{\tau,h}$ in order to estimate future parameters. Therefore, Step 2 can be reformulated as: $\forall\, i = 1, \ldots, N$:

$$\min_{\alpha_i} \frac{1}{2}\sum_{\tau=1}^{t-1}\left(\widehat{p_i^\tau} - \alpha_{i0}^\tau + \sum_{j=1}^{N}\alpha_{ij}^\tau d_j^\tau\right)^2 + \frac{1}{2}\sum_{\tau=t}^{T}\left(p_i^{\tau,h} - \alpha_{i0}^\tau + \sum_{j=1}^{N}\alpha_{ij}^\tau d_j^\tau\right)^2 \quad (12)$$

$$\begin{aligned}
\text{s.t} \qquad\qquad \alpha_{ij}^\tau &\geqslant \varepsilon \qquad \forall\, j = 1, \ldots, N, \tau = 1, \ldots, T \\
(1-\varepsilon)\alpha_{ii}^\tau &\geqslant \sum_{j\neq i}|\alpha_{ij}^\tau| \qquad \forall \tau = 1, \ldots, T \\
(1-\varepsilon)\alpha_{ii}^\tau &\geqslant \sum_{j\neq i}|\alpha_{ji}^\tau| \qquad \forall \tau = 1, \ldots, T \\
\|\alpha_i^\tau - \alpha_i^{\tau+1}\| &\leqslant \delta_i \|\alpha_i^\tau\| \qquad \forall \tau = 1, \ldots, T-1 \\
d_j = (d_j^1, \ldots, d_j^T) &\in \mathcal{BR}_j(d_{-j}, \alpha_j) \qquad \forall j = 1, \ldots, N
\end{aligned}$$

The differences with the previous model (equations (5) to (12)) lies in the fact that the estimation (equation (17)) is performed for all time periods up to T instead of only past time periods. Thus at each time period t, we seek to minimize the sum of two types of error: the squared error between the prices observed on the market, and the estimate given by the price function for all past periods 1 to $t-1$, and the squared error between the historical prices and the price function for periods t to T. Therefore, unlike in the previous model for Step 2, we need not find estimates of future price parameters as a function of past parameters. The solution to this problem is the set of optimal parameters for firm i for all time periods.

3.3 Step 3: Price-Setting for Future Periods

Steps 1 and 2 yield the estimates of the market-equilibrium demands, and of the price-demand parameters. Each firm can therefore use these estimates to set its prices for periods t to T, as well as guess its competitors' optimal pricing strategy, by plugging the estimates computed into the price function.

In this subsection, we now assume the position of one of the firms on the market, say firm i.

In fact, by taking the point of view of one of the firms in the market, say firm i, we can incorporate the fact that firm i knows more about itself than it does about its competitors. Indeed, it can observe its own past demand. Hence at period t, the demands $d_i^1, \ldots, d_i^{t-1}$ are data to firm i rather than unknown variables. We denote by $\widehat{d_i^1}, \ldots, \widehat{d_i^{t-1}}$ the observed values for firm i's demands. As a consequence, once Steps 1 and 2 are solved, Firm i can recompute a better estimate for its price-demand parameters and can also recompute its best response policy given that its past demands are data. The best-response problem for firm i thus becomes:

$$\max_{d_i^t,\ldots,d_i^T} \sum_{\tau=t}^{T} d_i^\tau \left(\alpha_{i0}^\tau - \sum_{j=1}^{N} \alpha_{ij}^\tau d_j^\tau \right) \tag{13}$$

$$\text{s.t} \qquad \sum_{\tau=t}^{T} d_i^\tau \leqslant C_i^\tau \tag{14}$$

$$d_j^\tau \geqslant 0 \qquad \forall\, \tau = t, \ldots, T \tag{15}$$

This best-response problem differs from that computed in Step 1 in 2 ways: first of all, the objective function (13) is the maximization of the revenue for future periods t through T; second, the constraint that the total demand does not exceed capacity has been replaced by constraint (14), where $C_i^\tau = C_i - \sum_{\tau=1}^{t-1} \widehat{d_i^\tau}$ is firm i's remaining capacity at time t: it states that the demand over future periods should not exceed the remaining capacity at t. The competitors 's demands remain the same as in Step 1, due to the asymmetry of information: indeed, they do not observe firm i's demand.

Taking this modification into account, the estimation problem for firm i becomes:

$$
\begin{aligned}
\min_{\alpha_i} \ & \frac{1}{2}\sum_{\tau=1}^{t-1}\left(\widehat{p_i^\tau} - \alpha_{i0}^\tau + \alpha_{ii}^\tau d_i^\tau + \sum_{j=1, j\neq i}^{N} \alpha_{ij}^\tau d_j^\tau\right)^2 \\
\text{s.t} \quad & \alpha_{ij}^\tau \geqslant \varepsilon && \forall\, i, j = 1, \ldots, N,\ \forall\, \tau = 1, \ldots, T \\
& (1-\varepsilon)\alpha_{ii}^\tau \geqslant \sum_{j\neq i} \alpha_{ij}^\tau, && \forall\, i = 1, \ldots, N,\ \forall \tau = 1, \ldots, T \\
& (1-\varepsilon)\alpha_{ii}^\tau \geqslant \sum_{j\neq i} \alpha_{ji}^\tau, && \forall\, i = 1, \ldots, N,\ \forall \tau = 1, \ldots, T \\
& \|\alpha_i^t - \alpha_i^{t+1}\| \leqslant \delta_i \|\alpha_i^t\|, && \forall\, i = 1, \ldots, N,\ \forall\, \tau = 1, \ldots, T-1 \\
& \alpha_i^\tau = \sum_{\theta=1}^{\tau} \omega^\theta \alpha_i^\theta, && \forall \tau = t+1, \ldots, T \\
& d_j \in \mathcal{BR}_j(\alpha_j, d_{-j}) && \forall j \neq i
\end{aligned} \tag{16}
$$

The difference between the above optimization problem and that solved in Step 2 for firm i lies in the fact that the demands for firm i are not those determined through the Nash equilibrium problem, but the observed demands $\widehat{d_i^\tau}$ $\tau = 1, \ldots, T$. On the other hand, the estimation problems of firm i's competitors are those of Step 2.

As a result, the set of Nash equilibrium demands, and the set of parameters simultaneously solving the N estimation problems are firm i's estimates of the best-response demands and optimal parameters, and can be used in order to compute its own, as well as its competitors' optimal pricing strategy for future periods. If d and α denote the solutions to Steps 1 and 2, as modified by firm i, then the set of optimal prices for future periods are: $\forall \tau = t, \ldots, T$, and $\forall j = 1, \ldots, N$:

$$p_j^\tau = \alpha_{j0}^\tau - \sum_{k=1}^{N} \alpha_{jk}^\tau d_k^\tau$$

In particular, firm i sets its price for period t to:

$$p_i^t = \alpha_{i0}^t - \sum_{j=1}^{N} \alpha_{ij}^t d_j^t$$

3.4 A Transient Learning Approach

At the beginning of each time period $t = 1, \ldots, T$, each firm solves the three-step dynamic pricing problem. Therefore, at each time period, it updates its estimate of the price functions of all firms, as well as setting its optimal prices for period t. It incorporates each time the additional information collected at the previous time period, in order to improve its estimates. On can therefore say that the firm learns the price-demand relationship as time goes on.

4. The Dynamic Pricing Problem as an MPEC

In this section, we reformulate the previous model using ideas from the theory of variational inequalities. We show that the Nash equilibrium problem in Step 1 can be reformulated as a Joint Variational Inequality. This Joint Variational Inequality can be interpreted as a user-equilibrium problem in the transportation setting. Therefore, we will exploit this analogy by using methods from the transportation field in order to solve the Joint Variational Inequality. (See Subsection 4.2 for more details).

4.1 Reformulation of Step 1

First of all, we state and prove a famous result: under some assumptions satisfied by our model, the optimization problem $\mathcal{BR}_i(d_{-i}, \alpha_i)$ is equivalent to a Variational Inequality.

PROPOSITION 4.1. *If the objective function $-\Pi_i$ of $\mathcal{BR}_i(d_{-i}, \alpha_i)$ is a continuously differentiable and convex function of the variable d_i and the feasible set $K_i = \{d_i \geqslant 0 \mid \sum_{\tau=1}^{T} d_i^\tau \leqslant C_i; \quad \alpha_{i0}^\tau - \sum_{j=1}^{N} \alpha_{ij}^\tau d_j^\tau \geqslant 0,\ \forall\, \tau = 1, \ldots, T\}$ is a closed and convex set, then the problem:*

$$\min_{d_i} -\Pi_i(d_i)$$
$$s.t \quad d_i \in K_i$$

is equivalent to the Variational Inequality denoted VI_i:

$$-\nabla\Pi_i(d_i)'(\tilde{d}_i - d_i) \geqslant 0, \quad \forall\, \tilde{d}_i \in K_i$$

PROOF. K_i is a closed and convex set due to the continuity and convexity of the functions that define its constraints.

Furthermore, the objective function:

$$-\Pi_i(d_i) = -\sum_{\tau=1}^{T} d_i^\tau \left(\alpha_{i0}^\tau - \sum_{k=1}^{N} \alpha_{ik}^\tau d_k^\tau \right)$$

is continuously differentiable. Its gradient is a vector of T components, each component being:

$$-\frac{\partial \Pi_i}{\partial d_i^\tau} = 2\alpha_{ii}^\tau d_i^\tau + \sum_{k\neq i} \alpha_{ik}^\tau d_k^\tau - \alpha_{i0}^\tau$$

Its Hessian matrix is a $T \times T$ diagonal matrix of coefficients: $2\alpha_{ii}^\tau$. Moreover, since the α_{ii}^τ are positive by assumption, the Hessian matrix of the objective function is positive definite, hence $-\Pi_i$ is convex. Therefore, the optimization problem $\mathcal{BR}_i(d_{-i}, \alpha_i)$ is equivalent to a variational inequality VI_i:

$$\sum_{\tau=1}^{T} \left(2\alpha_{ii}^\tau d_i^\tau + \sum_{j\neq i} \alpha_{ij}^\tau d_j^\tau - \alpha_{i0}^\tau \right) (\tilde{d}_i^\tau - d_i^\tau) \geqslant 0 \quad \forall \tilde{d}_i \in K_i$$

□

Furthermore, the following result holds:

PROPOSITION 4.2. *The N best response problems solved by each firm are equivalent to a single joint variational inequality denoted JVI:*

$$\sum_{j=1}^{N} \sum_{\tau=1}^{T} \left(2\alpha_{jj}^\tau d_j^\tau + \sum_{k\neq j} \alpha_{jk}^\tau d_k^\tau - \alpha_{j0}^\tau \right) (\tilde{d}_j^\tau - d_j^\tau) \geqslant 0$$

$$\forall \tilde{d} = (\tilde{d}_1, \ldots, \tilde{d}_N) \in K = K_1 \times \ldots \times K_N$$

PROOF. Let $d = (d_1, \ldots, d_N)$ be solution of the N Variational Inequalities: then $\forall\, i = 1, \ldots, N$, we have:

$$-\nabla \Pi_i(d_i)'(\tilde{d}_i - d_i) \geqslant 0 \; \forall\, \tilde{d}_i \in K_i$$

We can sum these N inequalities and obtain:

$$-\sum_{i=1}^{N} \nabla \Pi_i(d_i)'(\tilde{d}_i - d_i) \geqslant 0 \; \forall\, \tilde{d} \in K_1 \times \ldots \times K_N$$

Hence d is solution to the joint variational inequality JVI.

Conversely, if d is solution to JVI, then by choosing $\tilde{d}_j = d_j \;\forall\; j \neq i$, we get that d is also solution to VI_i.

Thus, the two formulations are equivalent. □

4.2 The Transportation Analogy

In this subsection, we focus on the case where the price-demand parameters are given (i.e. realized demand of competitors can be observed). Furthermore, we assume that the capacity constraint of the best-response problem of each firm is binding, i.e.:

$$\forall\, i = 1, \ldots, N, \quad \sum_{t=1}^{T} d_i^t = C_i$$

We show that under these assumptions, the best-response problem of a seller can be interpreted as a nonseparable user-equilibrium problem. Such a model is often used in the transportation field, as shown in Figure 1.

The analogy between this pricing problem and a Wardrop user-equilibrium problem is as follows: we have N origin-destination pairs (one for each seller). Each O-D pair is linked by T arcs, each arc corresponding to one time period. Seller i's capacity C_i becomes the total demand for travel from the origin to the destination point of i, i.e the number of travelers who want to travel from that origin to that destination. Moreover, the demand d_i^τ in time period τ of seller i becomes the flow on arc τ joining the O-D pair corresponding to i, that is the number of travelers who wish to travel on arc τ. t_i^τ corresponds to the travel time function on that arc.

The equivalent transportation problem corresponding to the maximization of the revenue of seller i over the selling horizon is the following: C_i selfish drivers try to minimize their travel time to go from an origin to a destination given the arc travel time functions shown in Figure 1. The solution to this problem will verify the following property (also called Wardrop second principle):

PROPOSITION 4.1 (Wardrop (1952)). *At equilibrium, all used arcs have equal and minimal travel times.*

Therefore, we can exploit this analogy by using methods from transportation problems in order to solve the joint variational inequality JVI. Indeed, the traffic equilibrium problem is a special class of optimization problems for which efficient and theoretically sound solution methods have been derived. In particular, we can use a relaxation scheme to solve the joint variational inequality. It is a special case of a general iterative scheme devised by Dafermos (1980).

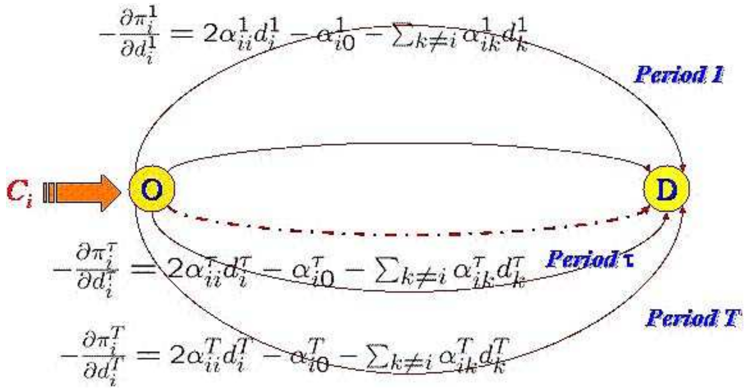


Figure 1. Best-response problem of Seller i with capacity C_i.

4.3 Steps 1 and 2 as a Mathematical Problem with Equilibrium Constraints

For fixed demands, each one of the N optimization problems solved in Step 2 has a convex objective function, and depends only on the parameters of the considered seller. Therefore, Step 2 is equivalent to a single optimization problem, obtained by summing the objective functions of the N problems solved in Step 2:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{\tau=1}^{T} \left(\hat{p_i^\tau} - \alpha_{i0}^\tau + \sum_{j=1}^{N} \alpha_{ij}^\tau d_j^\tau \right)^2 \tag{17}$$

$$\begin{aligned}
\text{s.t} \qquad \alpha_{ij}^\tau &\geqslant \varepsilon && \forall\, i, j = 1, \ldots, N,\ \forall \tau = 1, \ldots, T \\
(1-\varepsilon)\alpha_{ii}^\tau &\geqslant \sum_{j \neq i} \alpha_{ij}^\tau, && \forall\, i = 1, \ldots, N,\ \forall \tau = 1, \ldots, T \\
(1-\varepsilon)\alpha_{ii}^\tau &\geqslant \sum_{j \neq i} \alpha_{ji}^\tau, && \forall\, i = 1, \ldots, N,\ \forall \tau = 1, \ldots, T \\
\|\alpha_i^t - \alpha_i^{t+1}\| &\leqslant \delta_i \|\alpha_i^t\|, && \forall\, i = 1, \ldots, N,\ \forall\, \tau = 1, \ldots, T-1 \\
\alpha_i^\tau &= \sum_{\theta=1}^{\tau} \omega^\theta \alpha_i^\theta, && \forall \tau = t+1, \ldots, T
\end{aligned}$$

Furthermore, the solution to Step 1 is given as a function of the price-demand parameters, which are then estimated in Step 2. Therefore, Step 1 can be seen as the lower level of the optimization problem solved in Step 2. Hence the two steps can be formulated as a single problem involving two levels of optimization, the lower level being a joint variational inequality. Such a problem is called a Mathematical Program with Equilibrium Constraints

(MPEC):

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{\tau=1}^{T} \left(\widehat{p_i^{\tau}} - \alpha_i^{\tau} + \alpha_{ii}^{\tau} d_i^{\tau} + \sum_{j \neq i}^{N} \alpha_{ij}^{\tau} d_j^{\tau} \right)^2$$

$$\begin{aligned}
\text{s.t} \qquad & \alpha_{ij}^{\tau} \geqslant \varepsilon, && \forall\, i, j = 1, \ldots, N,\ \forall \tau = 1, \ldots, T \\
& (1-\varepsilon)\alpha_{ii}^{\tau} \geqslant \sum_{j \neq i} \alpha_{ij}^{\tau}, && \forall\, i = 1, \ldots, N,\ \forall \tau = 1, \ldots, T \\
& (1-\varepsilon)\alpha_{ii}^{\tau} \geqslant \sum_{j \neq i} \alpha_{ji}^{\tau}, && \forall\, i = 1, \ldots, N,\ \forall \tau = 1, \ldots, T \\
& \|\alpha_i^{t} - \alpha_i^{t+1}\| \leqslant \delta_i \|\alpha_i^{t}\|, && \forall\, i = 1, \ldots, N,\ \forall\, \tau = 1, \ldots, T-1 \\
& \alpha_i^{\tau} = \sum_{\theta=1}^{\tau} \omega^{\theta} \alpha_i^{\theta}, && \forall \tau = t+1, \ldots, T
\end{aligned}$$

$$\sum_{j=1}^{N} \sum_{\tau=1}^{T} \left(2\alpha_{jj}^{\tau} d_j^{\tau} - \alpha_{j0}^{\tau} + \sum_{k \neq j}^{N} \alpha_{jk}^{\tau} d_k^{\tau} \right) \left(\tilde{d}_j^{\tau} - d_j^{\tau} \right) \geqslant 0 \tag{18}$$

$$\forall\, \tilde{d} = (\tilde{d}_1, \ldots, \tilde{d}_N) \in K = K_1 \times \ldots \times K_N$$

where $K_i = \{d_i \geqslant 0 \mid \sum_{\tau=1}^{T} d_i^{\tau} \leqslant C_i; \quad \alpha_{i0}^{\tau} - \sum_{j=1}^{N} \alpha_{ij}^{\tau} d_j^{\tau} \geqslant 0,\ \forall\ \tau = 1, \ldots, T\}$.

The main difference between the above problem and equations (5) to (12) corresponding to Step 2 is that constraint (12) corresponding to the fact that the demands are the Nash equilibrium demands which were determined in Step 1, has been replaced by constraint (19) (i.e the joint variational inequality JVI). Therefore, instead of solving Step 1 separately, as a function of α, and then solving for α in Step 2, we can solve them jointly.

5. Computational Results

In this paper, we investigate two topics from a computational perspective. First, we would like to explore the market behavior under the proposed demand model. Second, we would like to test the joint dynamic pricing and learning approach.

5.1 The Demand Model

We study the demand model in a market consisting of 2 firms, and a selling horizon of 2 time periods, in order to gain some insights in the market equilibrium demands computed in the previous section. We carry out a sensitivity analysis of the Nash equilibrium demands, prices and revenues when the parameters and the capacities vary.

5.1.1 Methodology

We fix all but one of the coefficients of the price function of firm 1, and then compute the market equilibrium demands, prices and revenues for each firm and in each time period, as a function of the varying parameter α^1_{10}, α^1_{11} or α^1_{12}; furthermore, we repeat the process for various values of the capacities, in order to capture the effect of the capacities as well.

We start with the following set of values for the demand functions:

$$d^1_1 = 100 - 5p^1_1 + 3p^1_2 \qquad d^1_2 = 110 + 2.7p^1_1 - 4.5p^1_2$$
$$d^2_1 = 100 - 4.5p^2_1 + 2.7p^2_2 \qquad d^2_2 = 110 + 2.3p^2_1 - 4p^2_2$$

In other words, the price functions are approximately the following:

$$p^1_1 = 54.1667 - 0.3125d^1_1 - 0.2083d^1_2 \quad p^1_2 = 56.9444 - 0.1875d^1_1 - 0.3472d^1_2$$
$$p^2_1 = 59.1179 - 0.3393d^2_1 - 0.2290d^2_2 \quad p^2_2 = 61.4928 - 0.1951d^2_1 - 0.3817d^2_2$$

For each one of the parameters α^1_{10}, α^1_{11}, α^1_{12}, we plot the Nash equilibrium demands, prices and revenues for each time period and for each firm, as a function of this parameter.

5.1.2 Comparison of the Nash Equilibrium Demands, Prices and Revenues of the Firms

It is obvious that if their parameters are equal, then both firms will have equal demands, prices and revenues at equilibrium. We would like to compare the firms' equilibrium when we deviate from the symmetric case. The symmetric price functions are the following:

$$p_1 = 25 - 0.3125d_1 - 0.1875d_2$$
$$p_2 = 25 - 0.1875d_1 - 0.3125d_2$$

Furthermore, we consider 2 cases as far as capacities are concerned: both firms have tight capacity or both have unused capacity. We study asymmetric cases as far as the price functions are concerned: first, we study the Nash equilibrium when $\alpha_{11} > \alpha_{22}$, all other parameters being equal; then we study the case when $\alpha_{12} > \alpha_{21}$, all other parameters being equal. Finally, we focus on the Nash equilibrium for $\alpha_{10} > \alpha_{20}$. The observations are the following:

- $\alpha_{11} > \alpha_{22}$: firm 1's price is more sensitive to its own demand than firm 2's price.

  If the firms' capacities are tight at equilibrium, then their demands both equal their capacity, but firm 2 prices higher, and therefore gets a bigger revenue.

  If the firms have unused capacity at equilibrium, then firm 1's demand and price are both lower than firm 2's, resulting in a lower revenue.
- $\alpha_{12} > \alpha_{21}$: firm 1's price is more sensitive to firm 2's demand than firm 2's price.

  In case of tight capacity, then firm 1 prices lower than firm 2, and gets a lower revenue;

  In case of unused capacity remaining, then both firm 1's demand and price are lower than firm 2's, hence firm 1's revenue is lower.
- $\alpha_{10} > \alpha_{20}$:

  In the tight capacity setting, firm 1 prices higher and therefore gets a higher revenue.

  In the case of unused capacity remaining, then firm 1 has a higher demand and charges a higher price, and therefore gets a higher revenue.

5.1.3 Sensitivity Analysis on the Nash Equilibrium

The observations below correspond to the graphs displayed at the end of the paper.

- First of all, we note that a change in a parameter α^1_{ij} affects the equilibrium demands, prices and revenues in the first period only, when there remains unused capacity at equilibrium; otherwise, the equilibrium demands, prices and revenues in both time periods, for both firms depend on α^1_{ij}.
- Furthermore, firm 1 is more affected by a change in his own parameter than his competitor: the magnitude of the change in firm 1's demand, price, or revenue at time 1 for a given change in α^1_{1j} is greater than that of firm 2.
- *Effect of an increase in* α^1_{10}:

  Figure 2 (resp. 3, resp. 4) displays the evolution of the equilibrium demands (resp. prices, resp. revenues) of the firms in each time period as a function of α^1_{10}.

  The main effect is an increase in d^1_1, p^1_1 and thus π^1_1, as well as a decrease in these quantities for firm 2, regardless of the capacity. When capacity is tight for firm 1, d^2_1 decreases, p^2_1 increases, resulting in a decrease in π^2_1. When firm 2's capacity is tight only, then d^2_2 increases, p^2_2 decreases, and π^2_2 increases. Finally, if both firm's capacities are tight, then d^2_2, p^2_2 and therefore π^2_2 also increase.
- *Effect of an increase in* α^1_{11}*:*

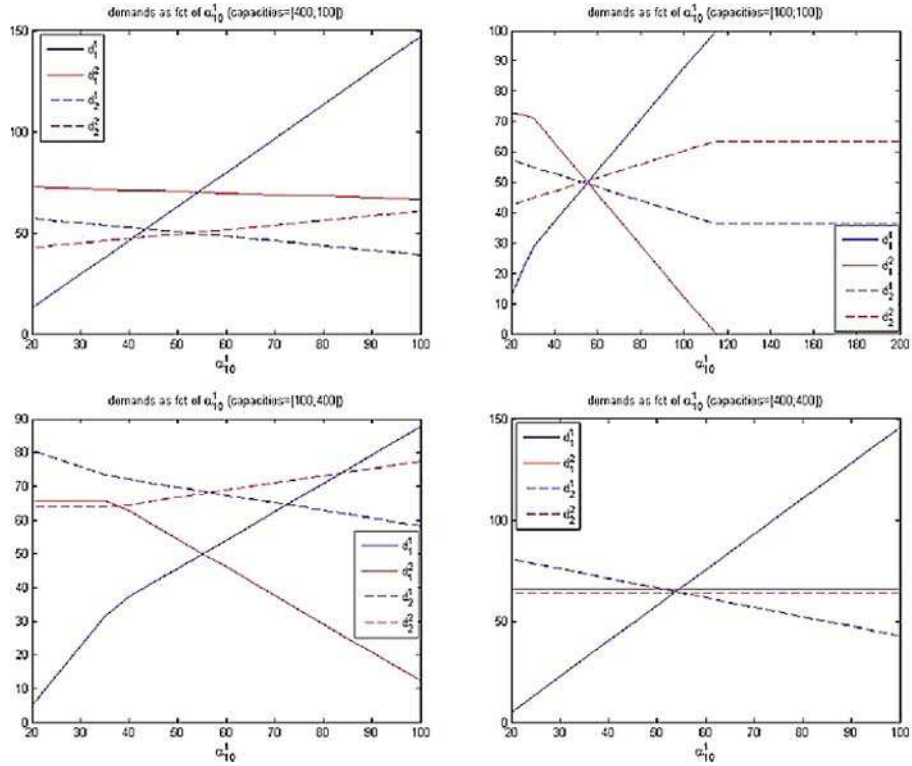


Figure 2. Nash equilibrium demands as a function of α_{10}^1.

Figure 2 (resp. 3, resp. 4) shows the evolution of the equilibrium demands (resp. prices, resp. revenues) of the firms in each time period as a function of α_{11}^1.

Whatever the capacities, d_1^1, p_1^1 and thus π_1^1 always decrease, whereas they increase for firm 2. On the other hand, if firm 1's capacity is tight, then we also observe that d_1^2 increases, p_1^2 decreases, and π_1^2 increases, whereas for firm 1's or firm 2's capacity tight, d_2^2 decreases, p_2^2 increases, and π_2^2 slightly decreases.

- *Effect of an increase in* α_{12}^1*:*

  Figure 2 (resp. 3, resp. 4) displays the evolution of the equilibrium demands (resp. prices, resp. revenues) of the firms in each time period as a function of α_{12}^1.

  Whatever the firms' capacities, we observe that firm 1's demand, price and revenue in period 1 decrease, whereas they increase for firm 2. When both firms' capacities are tight, then d_1^2 increases, p_1^2 decreases, and π_1^2 increases, whereas d_2^2, p_2^2 and π_2^2 all decrease. If firm 1's capacity is tight, not firm 2's,

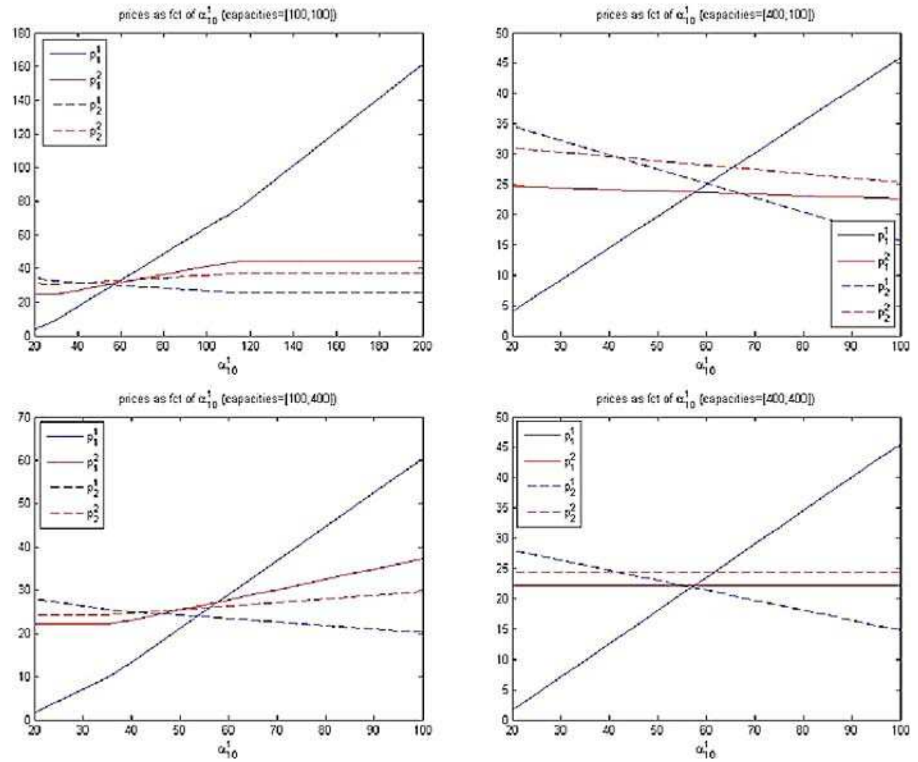


Figure 3. Nash equilibrium prices as a function of α_{10}^1.

then the demand, price and revenue for firm 1 in period 2 all increase; d_2^2 still decreases, but p_2^2 increases, resulting in a decrease in revenue. Finally, if firm 2's capacity is tight only, then d_1^2 still increases, but p_1^2 now decreases, and the corresponding revenue increases, whereas all three functions decrease for firm 2.

5.2 The Joint Dynamic Pricing and Learning Approach

5.2.1 Methodology

At each time period of the approach, we are solving a least squares problem in which we fit each data point $(p_i^\tau, d_1^\tau, \ldots, d_N^\tau)$ with coefficients $\alpha_{i0}^\tau, \alpha_{i1}^\tau, \ldots, \alpha_{iN}^\tau$. Hence to estimate $N + 1$ coefficients, we only rely on a single data point, and the problem is clearly underdetermined. In order to solve this issue, we assume that a certain number of price observations are available to each firm at each time period. For instance these might be price estimations from pre-sale data, or historical prices, etc. At each time period, the firm therefore has $Y > N + 1$ observed prices.

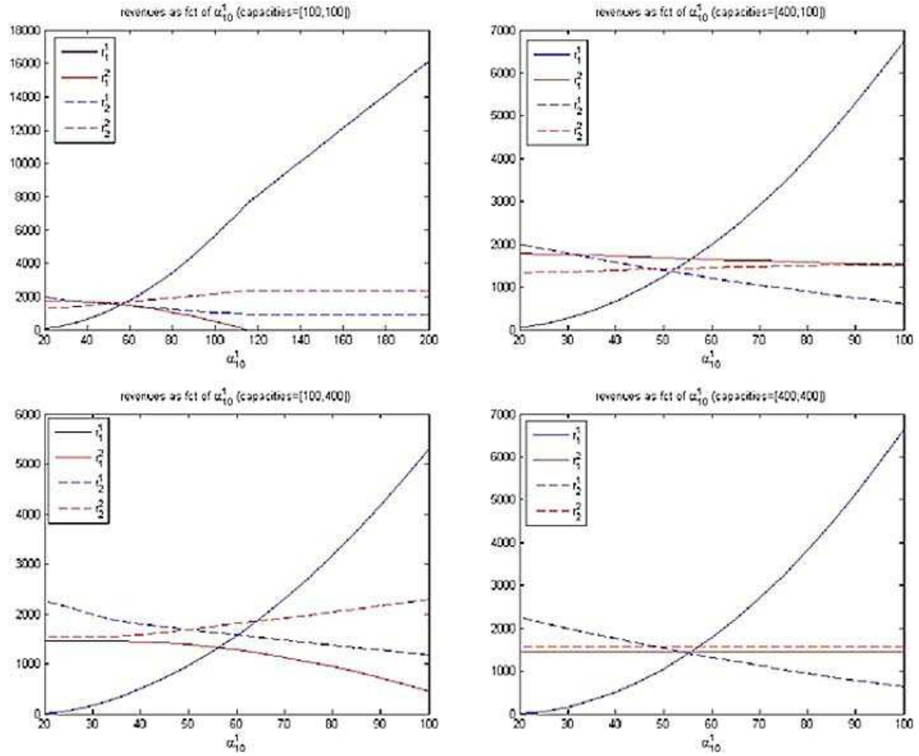


Figure 4. Nash equilibrium revenues as a function of α^1_{10}.

Furthermore, in order to assess the quality of the model, we assume that we know the true model of demand, i.e true values for the parameters β. This enables us to compute the corresponding demands and prices. The price observations are also obtained through simulation from using these parameters. We can then compare the parameters, prices and demands computed at each time period of the approach with the true values.
The input of the model are the following:

- N: number of firms in the market;
- T: number of time periods of the selling horizon;
- Y: number of price observations per time period;
- σ_p: standard deviation of the observed prices ;
- C_i: capacities of the firms ;
- α: true parameters (or equivalently β);
- δ: volatility of the parameters.

The computations were performed using Gauss-Newton method on a Unix workstation with 1GB of RAM and a processor of 2.4 GHz.

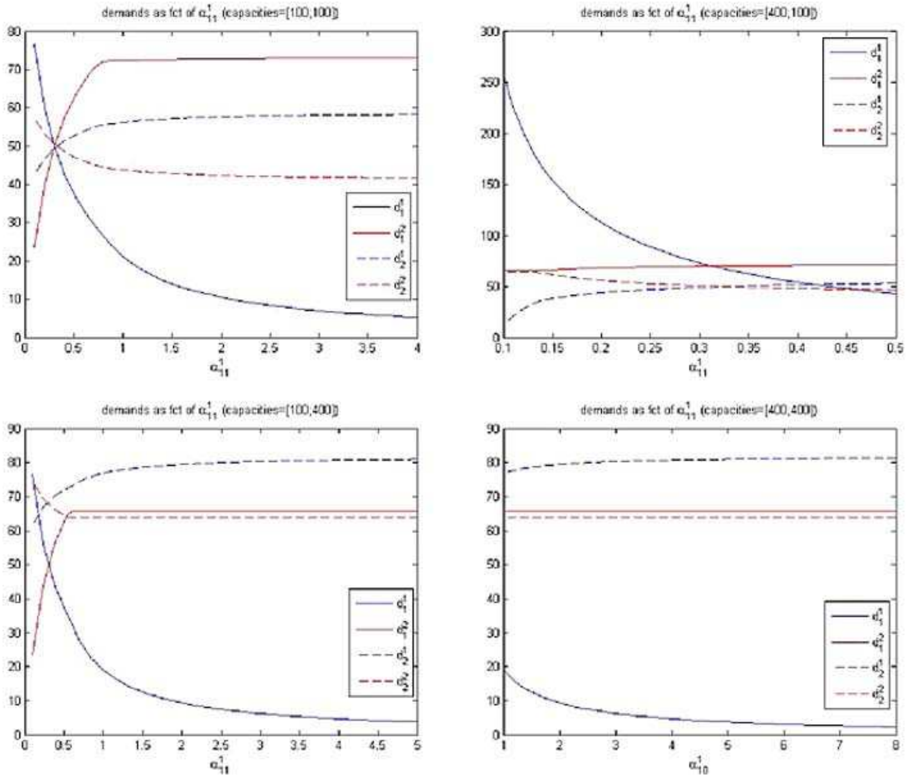


Figure 5. Nash equilibrium demands as a function of α_{11}^1.

In this section, we show the estimates of the parameters, demands, prices and revenues computed by the approach. We focus on the 2firm-2-period example. We assume $Y = 30$ price observations available at each time period, with standard deviation $\sigma_p = 1$. The firms' capacities are $C_1 = 100$, $C_2 = 100$. We assume the true model of demand to be:

$$\alpha_{ij}^{\tau} = \bar{\alpha}_{ij}^{\tau} + \epsilon_{ij}^{\tau}$$

where ϵ_{ij}^{τ} is chosen to be gaussian with mean 0 and standard deviation $\sigma_{i0}^{\tau} = 0.5$ and $\sigma_{ij}^{\tau} = 0.05$.

The values $\bar{\alpha}_{ij}^{t}$ are such that:

$$p_1^1 = 54.1667 - 0.3125 d_1^1 - 0.2083 d_2^1 \quad p_1^2 = 59.1179 - 0.3393 d_1^2 - 0.2290 d_2^2$$
$$p_2^1 = 56.9444 - 0.1875 d_1^1 - 0.3472 d_2^1 \quad p_2^2 = 61.4928 - 0.1951 d_1^2 - 0.3817 d_2^2$$

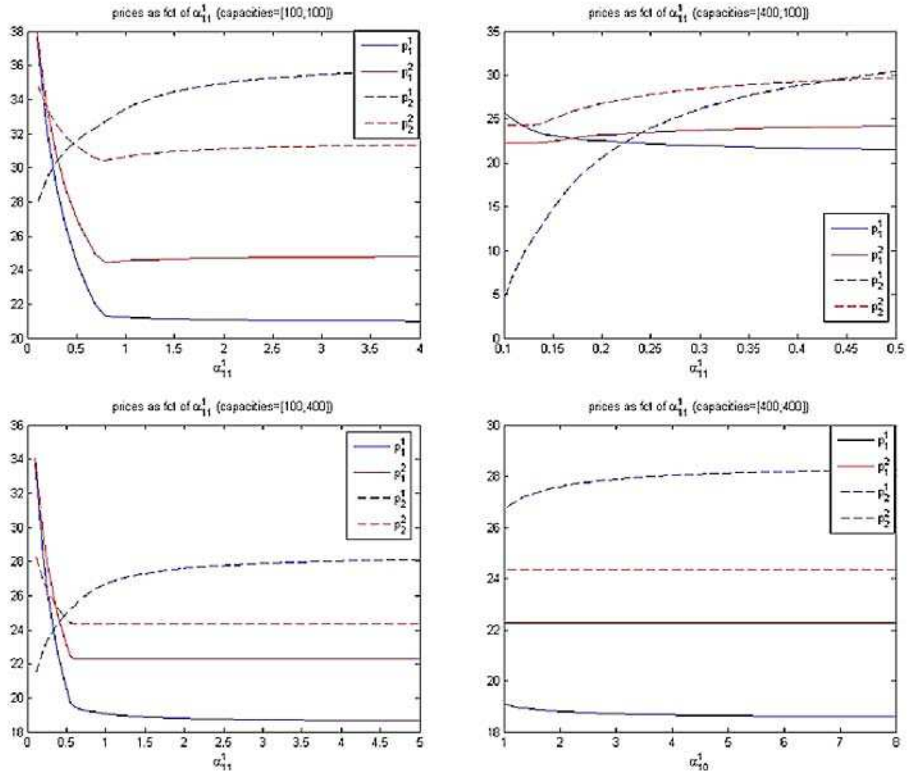


Figure 6. Nash equilibrium prices as a function of α^1_{11}.

We ran 1,000 simulations, and computed the optimal demands, parameters, and prices for each simulation, to obtain their average and standard deviation. We compared these to the average and standard deviation on the true demands, parameters and prices.

The results are the following:

- the average parameters obtained through simulation differ from the true parameters by 0.05%; the matrix of the variances of all the simulated parameters is:

$$\sigma(\alpha^1) = \begin{pmatrix} 0.2169 & 0.0022 & 0.0022 \\ 0.1969 & 0.0022 & 0.0028 \end{pmatrix}$$

$$\sigma(\alpha^2) = \begin{pmatrix} 0.2387 & 0.0023 & 0.0021 \\ 0.2393 & 0.0023 & 0.0023 \end{pmatrix}$$

In other words, the standard deviation of the constant parameters is of the order of 0.4, and that of the other parameters is order of 0.04.

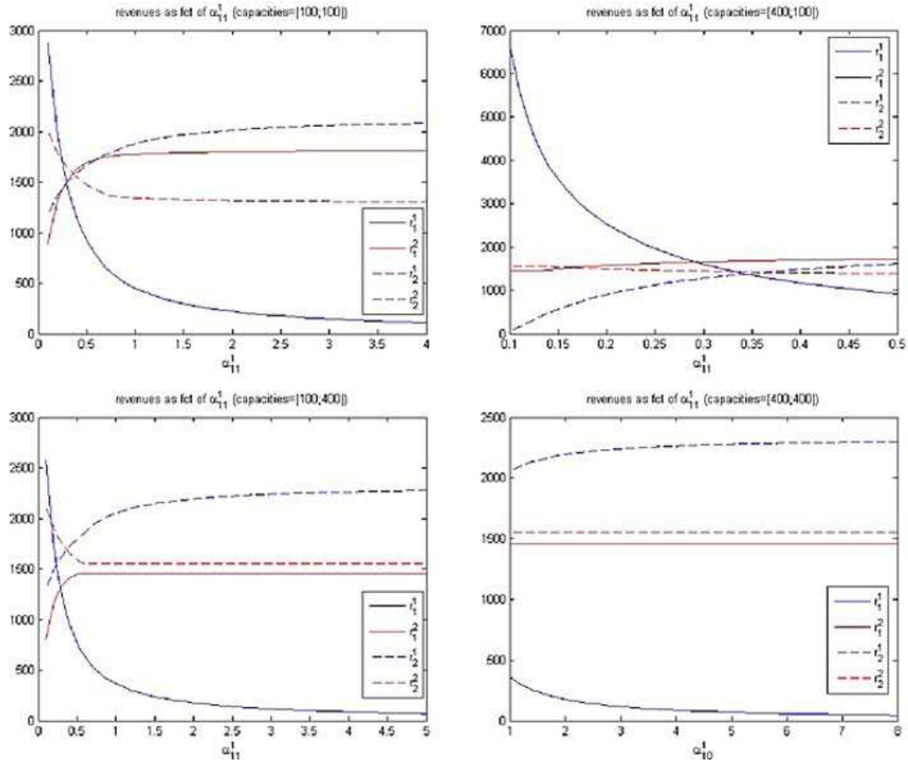


Figure 7. Nash equilibrium revenues as a function of α_{11}^1.

- The average demands obtained through simulation differ from the true demands by 0.56%; the matrix of standard deviation of the demands is:

$$\begin{pmatrix} \sigma(d_1^1) \; \sigma(d_1^2) \\ \sigma(d_2^1) \; \sigma(d_2^2) \end{pmatrix} = \begin{pmatrix} 6.77 \; 6.77 \\ 6.43 \; 6.23 \end{pmatrix}$$

- the average prices of the simulation differ from the true prices by 0.031%, and the matrix of the standard deviation of the prices is:

$$\begin{pmatrix} \sigma(p_1^1) \; \sigma(p_1^2) \\ \sigma(p_2^1) \; \sigma(p_2^2) \end{pmatrix} = \begin{pmatrix} 2.95 \; 2.87 \\ 2.87 \; 2.57 \end{pmatrix}$$

6. Application: 2-Firm 2-Period Dynamic Pricing Problem

We focus on the case of 2 firms and 2 pricing periods, in order to gain additional insights on the model and provide analytical evidence on the

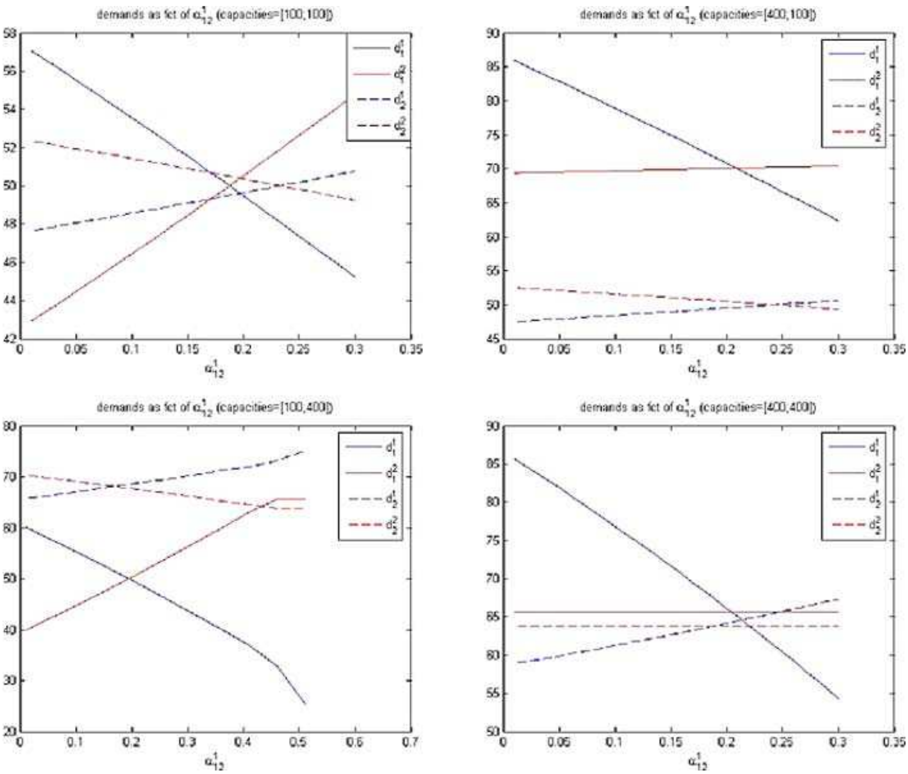


Figure 8. Nash equilibrium demands as a function of α_{12}^1.

computational results. Indeed, in this simple case, we can derive closed form solutions to the Nash equilibrium and estimation problem.

6.1 Solution to the Nash Equilibrium Problem

6.1.1 Best Response Problem of Each Firm

For the more general T-period problem, firm 1 solves the following best-response problem: for fixed demand of firm 2: $\mathcal{BR}_1(d_2, \alpha_1)$:

$$\begin{aligned} \max_{(d_1^1,\ldots,d_1^T)} \quad & \textstyle\sum_{t=1}^T d_1^t(\alpha_{10}^t - \alpha_{11}^t d_1^t - \alpha_{12}^t d_2^t) \\ \text{s.t.} \quad & \textstyle\sum_{t=1}^T d_1^t \leqslant C_1 \\ & d_1^t \geqslant 0 \end{aligned}$$

We now establish the following optimality conditions:

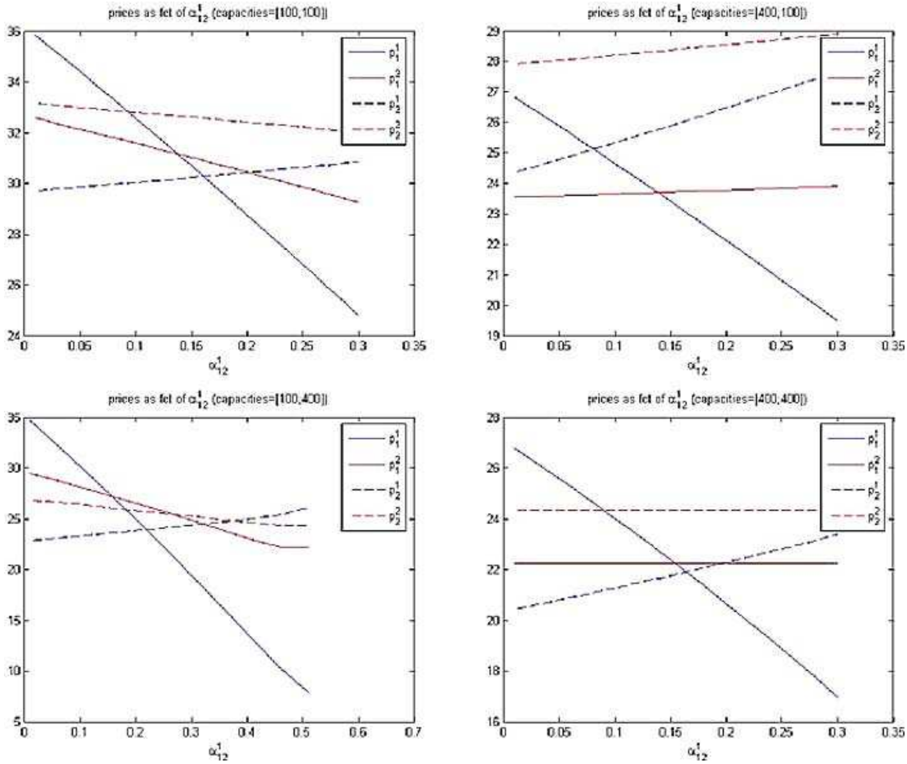


Figure 9. Nash equilibrium prices as a function of α^1_{12}.

PROPOSITION 6.1. *The Karush-Kuhn-Tucker conditions are necessary and sufficient for optimality of* $\mathcal{BR}_1(d_2, \alpha_1)$.

PROOF. The gradients of the inequality constraints are linearly independent, establishing necessity.

Furthermore, $-\Pi_1$ is convex, and the functions defining the inequality constraints are linear, establishing sufficiency. □

The KKT conditions can be written as:

$$\frac{\partial \Pi_1}{\partial d_1^t} - \lambda_1 \leqslant 0 \tag{19}$$

$$d_1^t \left(\frac{\partial \Pi_1}{\partial d_1^t} - \lambda_1 \right) = 0 \tag{20}$$

$$\sum_{t=1}^{T} d_1^t \leqslant C_1 \tag{21}$$

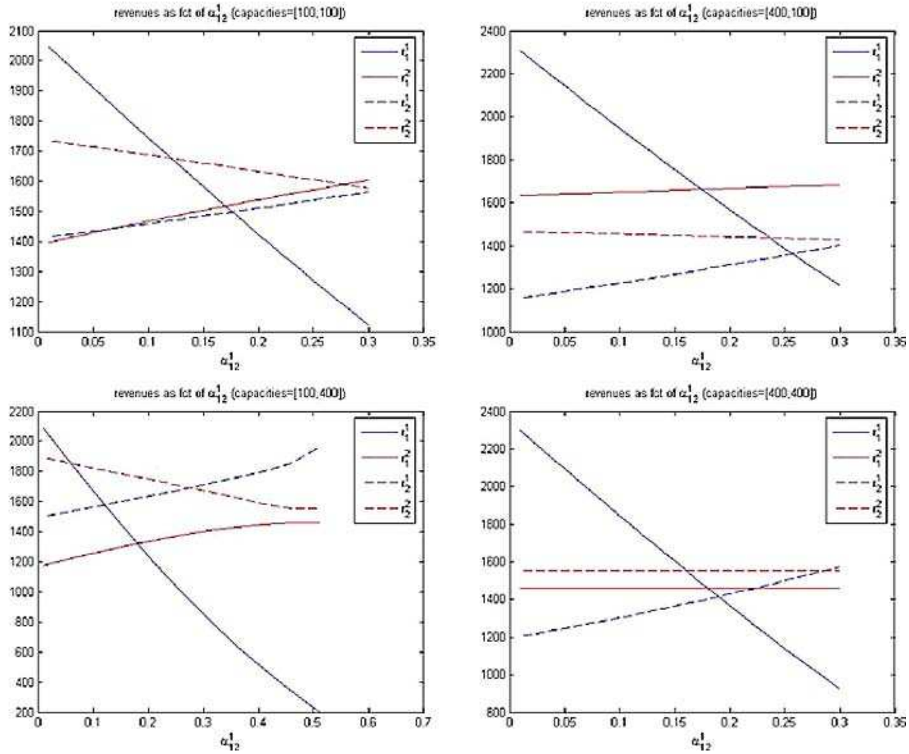


Figure 10. Nash equilibrium revenues as a function of α_{12}^{1}.

$$\lambda_1 \left(\sum_{t=1}^{T} d_1^t \leqslant C_1 \right) = 0 \tag{22}$$

$$d_1^t \geqslant 0 \ \forall t = 1, \ldots, T \tag{23}$$

$$\lambda_1 \geqslant 0 \tag{24}$$

where λ_1 is the Lagrange multiplier corresponding to the capacity constraint. Constraints (21) and (22) are the first order optimality conditions: note that if $d_1^t > 0$, then constraint (22) has to be satisfied with equality: $\frac{\partial \Pi_1}{\partial d_1^t} - \lambda_1 = 0$. Constraints (21) and (22) can be explicitly written as:

$$\alpha_{10}^t - 2\alpha_{11}^t d_1^t - \alpha_{12}^t d_2^t - \lambda_1 \leqslant 0$$
$$d_1^t(\alpha_{10}^t - 2\alpha_{11}^t d_1^t - \alpha_{12}^t d_2^t) = 0$$

Constraint (23) is just the capacity constraint. Constraint (24) is the complementary slackness condition: the multiplier λ_1 is equal to zero unless the capacity constraint is tight.

Hence two cases arise: either the capacity constraint is tight, i.e all capacity is used at the optimum, or there remains unused capacity at the optimum. The following proposition gives the characterization of firm 1's best-response demands.

PROPOSITION 6.2.

1. *If* $\sum_{t=1}^{T} \frac{\alpha_{10}^t - \alpha_{12}^t d_2^t}{2\alpha_{11}^t} < C_1$, *then firm 1's best response demands are equal to:* $\forall\, t$ *such that* $d_1^t > 0$*:*

$$d_1^t = \frac{\alpha_{10}^t - \alpha_{12}^t d_2^t}{2\alpha_{11}^t} \tag{25}$$

2. *Else, if* $\sum_{t=1}^{T} \frac{\alpha_{10}^t - \alpha_{12}^t d_2^t}{2\alpha_{11}^t} \geqslant C_1$, *then firm 1's best-response demands are:* $\forall\, t$ *such that* $d_1^t > 0$*:*

$$d_1^t = \frac{\alpha_{10}^t - \alpha_{12}^t d_2^t}{2\alpha_{11}^t} - \frac{\lambda_1}{2\alpha_{11}^t} \tag{26}$$

where $\lambda_1 = (\sum_{t=1}^{T} \frac{1}{2\alpha_{11}^t})^{-1} \left(\sum_{t=1}^{T} \frac{\alpha_{10}^t - \alpha_{12}^t d_2^t}{2\alpha_{11}^t} - C_1 \right)$

6.1.2 Computing the Nash Equilibrium

The Nash equilibrium demands are the demands which simultaneously solve both firms' best-response problems. Three cases arise at equilibrium: both firms have unused capacity, exactly one firm has unused capacity, or both firms have tight capacity. We now characterize the Nash equilibrium of the 2-firm-2-period problem in each case:

PROPOSITION 6.3. *If both firms have unused capacity at equilibrium, then the Nash equilibrium demands are:*

$$\begin{pmatrix} d_1^1(\alpha) \\ d_2^1(\alpha) \\ d_1^2(\alpha) \\ d_2^2(\alpha) \end{pmatrix} = \begin{pmatrix} 2\alpha_{22}^1/\Delta^1 & -\alpha_{12}^1/\Delta^1 & 0 & 0 \\ -\alpha_{21}^1/\Delta^1 & 2\alpha_{11}^1/\Delta^1 & 0 & 0 \\ 0 & 0 & 2\alpha_{22}^2/\Delta^2 & -\alpha_{12}^2/\Delta^2 \\ 0 & 0 & -\alpha_{21}^2/\Delta^2 & 2\alpha_{11}^2/\Delta^2 \end{pmatrix} \begin{pmatrix} \alpha_{10}^1 \\ \alpha_{20}^1 \\ \alpha_{10}^2 \\ \alpha_{20}^2 \end{pmatrix} \tag{27}$$

where $\Delta^t = 4\alpha^t_{11}\alpha^t_{22} - \alpha^t_{12}\alpha^t_{21}$. *This holds for parameters* α *such that:*

$$\frac{2\alpha^1_{22}\alpha^1_{10} - \alpha^1_{12}\alpha^1_{20}}{\Delta^1} + \frac{2\alpha^2_{22}\alpha^2_{10} - \alpha^2_{12}\alpha^2_{20}}{\Delta^2} \leqslant C_1$$
$$\frac{2\alpha^1_{11}\alpha^1_{20} - \alpha^1_{21}\alpha^1_{10}}{\Delta^1} + \frac{2\alpha^2_{11}\alpha^2_{20} - \alpha^2_{21}\alpha^2_{10}}{\Delta^2} \leqslant C_2$$

We make the following observations concerning the influence of the parameters on the Nash equilibrium demands:

- Nash equilibrium demand of firm 1 as a function of α_{11}:
  $d^t_1(\alpha)$ is decreasing in α^t_{11}, namely when firm 1's price is more sensitive to its own demand, then its demand increases. This is in agreement with Figure 5.
- Nash equilibrium demand of firm 1 as a function of α_{22}:
  If $\alpha^t_{12} > 0$ then d^t_1 is decreasing in α^t_{22}. If the sensitivity of firm 1's demand to firm 2's price is positive (i.e when firm 2's price is greater, firm 1's demand also is), then $d^t_1(\alpha)$ decreases as firm 2's demand is more sensitive to changes in its own prices.
  If $\alpha^t_{12} < 0$ then $d^t_1(\alpha)$ is increasing in α^t_{22}.
- Nash equilibrium demand of firm 1 as a function of α_{12}:
  $d^t_1(\alpha)$ is decreasing in α^t_{12} , as illustrated in Figure 8. When firm 1's demand is more sensitive to firm 2's price, then its demand at equilibrium is lower.
- Nash equilibrium demand of firm 1 as a function of α_{21}:
  $d^t_1(\alpha)$ is increasing in α^t_{21}: as firm 2's demand becomes more sensitive to firm 1's price, then firm 1's equilibrium demand is higher.
- Nash equilibrium demand of firm 1 as a function of α_{10}:
  d^1_1 is increasing in α^t_{10}. This corresponds to Figure 2.
- Nash equilibrium demand of firm 1 as a function of α_{20}:
  If $\alpha^t_{12} > 0$, then d^t_1 is decreasing in α^t_{20}.

PROPOSITION 6.4. *If both firms' capacities are tight at equilibrium, then let* $\bar{\alpha}_{ij} = \alpha^1_{ij} + \alpha^2_{ij}$ *and* $\Delta = 4\bar{\alpha}_{11}\bar{\alpha}_{22} - \bar{\alpha}_{12}\bar{\alpha}_{21}$. *The Nash equilibrium demands are given by:*

$$\begin{pmatrix} d^1_1(\alpha) \\ d^1_2(\alpha) \\ d^2_1(\alpha) \\ d^2_2(\alpha) \end{pmatrix} = \frac{1}{\Delta} \begin{pmatrix} 2\bar{\alpha}_{22} - \frac{\alpha^2_{21}\bar{\alpha}_{12}}{2\bar{\alpha}_{11}} & -\alpha^1_{12} & -\frac{\alpha^2_{21}\bar{\alpha}_{12}}{2\bar{\alpha}_{11}} & \alpha^2_{12} \\ -\alpha^1_{21} & 2\bar{\alpha}_{11} - \frac{\alpha^2_{12}\bar{\alpha}_{21}}{2\bar{\alpha}_{22}} & \alpha^2_{21} & -\frac{\alpha^2_{12}\bar{\alpha}_{21}}{2\bar{\alpha}_{22}} \\ -\frac{\alpha^1_{21}\bar{\alpha}_{12}}{2\bar{\alpha}_{11}} & \alpha^1_{12} & 2\bar{\alpha}_{22} - \frac{\alpha^1_{21}\bar{\alpha}_{12}}{2\bar{\alpha}_{11}} & -\alpha^2_{12} \\ \alpha^1_{21} & -\frac{\alpha^1_{12}\bar{\alpha}_{21}}{2\bar{\alpha}_{22}} & -\alpha^2_{21} & 2\bar{\alpha}_{11} - \frac{\alpha^1_{12}\bar{\alpha}_{21}}{2\bar{\alpha}_{22}} \end{pmatrix}$$

$$\times \begin{pmatrix} 2\alpha_{11}^2 C_1 + \alpha_{10}^1 - \alpha_{10}^2 \\ 2\alpha_{22}^2 C_2 + \alpha_{20}^1 - \alpha_{20}^2 \\ 2\alpha_{11}^1 C_1 + \alpha_{10}^2 - \alpha_{10}^1 \\ 2\alpha_{22}^1 C_2 + \alpha_{20}^2 - \alpha_{20}^1 \end{pmatrix} \tag{28}$$

This holds for parameters α *such that:*

$$\frac{2\alpha_{22}^1\alpha_{10}^1 - \alpha_{12}^1\alpha_{20}^1}{\Delta^1} + \frac{2\alpha_{22}^2\alpha_{10}^2 - \alpha_{12}^2\alpha_{20}^2}{\Delta^2} \geqslant C_1$$
$$\frac{2\alpha_{11}^1\alpha_{20}^1 - \alpha_{21}^1\alpha_{10}^1}{\Delta^1} + \frac{2\alpha_{11}^2\alpha_{20}^2 - \alpha_{21}^2\alpha_{10}^2}{\Delta^2} \geqslant C_2$$

Hence, when the capacity constraint of both firms is tight, the Nash equilibrium demand of each firm in one time period depends on the parameters of the firms in both time periods, as well as on the capacity of both firms.

As both firms have tight capacity, we know that $d_1^2(\alpha) = C_1 - d_1^1(\alpha)$ and $d_2^2(\alpha) = C_2 - d_2^1(\alpha)$. As a result, the Nash equilibrium is determined by the values of the demands in period 1, and we can write a simpler system solved by d_1^1 and d_2^1.

Let Δ, Δ^1 and Δ^2 be as defined above, and let:

$$[A^t] = \begin{pmatrix} 2\alpha_{11}^t & \alpha_{12}^t \\ \alpha_{21}^t & 2\alpha_{22}^t \end{pmatrix}$$

Finally, let $\bar{d}_1$ and $\bar{d}_2$ be the Nash equilibrium demands in the unconstrained case. Then in the case where both firms have tight capacity, the demands in period 1 at equilibrium are such that:

$$\begin{pmatrix} d_1^1(\alpha) \\ d_2^1(\alpha) \end{pmatrix} = \frac{\Delta^1}{\Delta}\begin{pmatrix} \bar{d}_1^1 \\ \bar{d}_1^2 \end{pmatrix} - \frac{\Delta^2}{\Delta}\begin{pmatrix} \bar{d}_2^1 \\ \bar{d}_2^2 \end{pmatrix} + \frac{\Delta^1}{\Delta}\Bigg[[A^1]^{-1}[A^2]\begin{pmatrix} C_1 \\ C_2 \end{pmatrix}$$
$$-[A^1]^{-1}\begin{pmatrix} \alpha_{10}^2 \\ \alpha_{20}^2 \end{pmatrix}\Bigg] + \frac{\Delta^2}{\Delta}\Bigg[\begin{pmatrix} C_1 \\ C_2 \end{pmatrix} - [A^2]^{-1}\begin{pmatrix} \alpha_{10}^1 \\ \alpha_{20}^1 \end{pmatrix}\Bigg]$$

Hence the Nash equilibrium in this case can be expressed as a function of the unconstrained Nash.

Similarly, we establish the following in the last case:

PROPOSITION 6.5. *If exactly one firm has unused capacity at equilibrium, say firm 2 without loss of generality, then the Nash equilibrium demands are:*

$$d_1^1(\alpha) = \left(\frac{\Delta^1}{2\alpha_{22}^1} + \frac{\Delta^2}{2\alpha_{22}^2}\right)^{-1}\left(\frac{\Delta^1}{2\alpha_{22}^1}\bar{d}_1{}^1 - \frac{\Delta^2}{2\alpha_{22}^2}\bar{d}_2{}^2 + \frac{\Delta^2}{2\alpha_{22}^2}C_1\right) \tag{29}$$

$$d_1^2(\alpha) = \left(\frac{\Delta^1}{2\alpha_{22}^1} + \frac{\Delta^2}{2\alpha_{22}^2}\right)^{-1} \left(\frac{\Delta^2}{2\alpha_{22}^2}\bar{d_1}^2 - \frac{\Delta^1}{2\alpha_{22}^1}\bar{d_2}^1 + \frac{\Delta^1}{2\alpha_{22}^1}C_1\right)$$

$$d_2^1(\alpha) = \frac{\alpha_{20}^1}{2\alpha_{11}^1} - \frac{\alpha_{21}^1}{2\alpha_{22}^1}\left(\frac{\Delta^1}{2\alpha_{22}^1} + \frac{\Delta^2}{2\alpha_{22}^2}\right)^{-1} \left(\frac{\Delta^1}{2\alpha_{22}^1}\bar{d_1}^1 - \frac{\Delta^2}{2\alpha_{22}^2}\bar{d_2}^2 + \frac{\Delta^2}{2\alpha_{22}^2}C_1\right)$$

$$d_2^2(\alpha) = \frac{\alpha_{20}^2}{2\alpha_{11}^2} - \frac{\alpha_{21}^2}{2\alpha_{22}^2}\left(\frac{\Delta^1}{2\alpha_{22}^1} + \frac{\Delta^2}{2\alpha_{22}^2}\right)^{-1} \left(\frac{\Delta^2}{2\alpha_{22}^2}\bar{d_1}^2 - \frac{\Delta^1}{2\alpha_{22}^1}\bar{d_2}^1 + \frac{\Delta^1}{2\alpha_{22}^1}C_1\right)$$

where $\bar{d_1}^t$, $\bar{d_2}^t$ *denote the solution to the market equilibrium in the unconstrained case. The set of parameters* α *for which this holds is:*

$$\frac{2\alpha_{22}^1\alpha_{10}^1 - \alpha_{12}^1\alpha_{20}^1}{\Delta^1} + \frac{2\alpha_{22}^2\alpha_{10}^2 - \alpha_{12}^2\alpha_{20}^2}{\Delta^2} \geqslant C_1$$

$$\frac{2\alpha_{11}^1\alpha_{20}^1 - \alpha_{21}^1\alpha_{10}^1}{\Delta^1} + \frac{2\alpha_{11}^2\alpha_{20}^2 - \alpha_{21}^2\alpha_{10}^2}{\Delta^2} \leqslant C_2$$

6.2 Solution to the Estimation Problem

Since we derived closed-form solutions for the Nash equilibrium demands, as a function of the parameters α, the MPEC is therefore equivalent to a single-level optimization problem in variable α. The set of feasible parameters α is partitioned into four regions, each region corresponding to one of the four Nash equilibria computed above (both firms have tight capacity, both have unused capacity, firm 1 has tight capacity only, firm 2 has tight capacity only).

For the sake of tractability, we relax the constraint of slowly varying parameters and consider the following simplifying assumption:

The matrix of price-demand parameters is such that:

$$\begin{pmatrix} \alpha_{11}^t & \alpha_{12}^t \\ \alpha_{22}^t & \alpha_{21}^t \end{pmatrix} = \begin{pmatrix} \alpha^t & 1 \\ 1 & \alpha^t \end{pmatrix}$$

In this simpler case, we derive the single-level problem which is equivalent to the MPEC.

PROPOSITION 6.6. *The objective function of the estimation problem is the following:*

$$\frac{1}{2}\sum_{t=1}^{T}\left(\widehat{p_1^t} - \alpha_{10}^t + \alpha^t\frac{\alpha_{10}^t + \alpha_{20}^t}{2\alpha_{10}^t + 1}\right)^2 + \left(\widehat{p_2^t} - \alpha_{20}^t + \alpha^t\frac{\alpha_{10}^t + \alpha_{20}^t}{2\alpha_{10}^t + 1}\right)^2$$

if the parameters α verify the following constraints:

$$\sum_{t=1}^{T} \frac{2\alpha^t \alpha_{10}^t - \alpha_{20}^t}{4(\alpha^t)^2 - 1} \leqslant C_1$$

$$\sum_{t=1}^{T} \frac{2\alpha^t \alpha_{20}^t - \alpha_{10}^t}{4(\alpha^t)^2 - 1} \leqslant C_2$$

$$\frac{1}{2}\Big(\widehat{p_1^1} - \alpha_{10}^1 + 2\alpha^1 \frac{\alpha^1 + \alpha^2 - 1}{4(\alpha^1 + \alpha^2)^2 + 1}\Big(2\big(\alpha^1 + \alpha^2 + 1\big)\big(C_1 + C_2\big) + \alpha_{10}^1 - \alpha_{10}^2$$
$$+\alpha_{20}^1 - \alpha_{20}^2\Big)\Big)^2 + \ldots + \frac{1}{2}\Big(\widehat{p_2^2} - \alpha_{20}^2 + 2\alpha^2 \frac{\alpha^1 + \alpha^2 - 1}{4(\alpha^1 + \alpha^2)^2 + 1}$$
$$\times\Big(2\big(\alpha^1 + \alpha^2 + 1\big)\big(C_1 + C_2\big) + \alpha_{20}^2 - \alpha_{20}^1 + \alpha_{10}^2 - \alpha_{10}^1\Big)\Big)^2$$

if the parameters α verify the following constraints:

$$\frac{2\alpha^1 \alpha_{10}^1 - \alpha_{20}^1}{4(\alpha^1)^2 - 1} + \frac{2\alpha^2 \alpha_{10}^2 - \alpha_{20}^2}{4(\alpha^2)^2 - 1} \geqslant C_1$$

$$\frac{2\alpha^1 \alpha_{20}^1 - \alpha_{10}^1}{4(\alpha^1)^2 - 1} + \frac{2\alpha^2 \alpha_{20}^2 - \alpha_{10}^2}{4(\alpha^2)^2 - 1} \geqslant C_2$$

$$\frac{1}{2}\Big(\widehat{p_1^1} - \alpha_{10}^1 + \big(\alpha^1 - 1/2\big)\Big(2\alpha^1 - \frac{1}{2\alpha^1} + 2\alpha^2 - \frac{1}{2\alpha^2}\Big)^{-1}\Big(\Big(\alpha_{10}^1 - \frac{\alpha_{20}^1}{2\alpha^1}\Big)$$
$$-\Big(\alpha_{20}^2 - \frac{\alpha_{10}^2}{2\alpha^2}\Big) + \Big(2\alpha^2 - \frac{1}{2\alpha^2}\Big)C_1\Big)\Big)^2 + \ldots$$
$$+\frac{1}{2}\Big(\widehat{p_2^2} - 1/2\Big(2\alpha^1 - \frac{1}{2\alpha^1} + 2\alpha^2 - \frac{1}{2\alpha^2}\Big)^{-1}\Big(\Big(\alpha_{10}^2 - \frac{\alpha_{20}^2}{2\alpha^2}\Big)$$
$$-\Big(\alpha_{20}^1 - \frac{\alpha_{10}^1}{2\alpha^1}\Big) + \Big(2\alpha^1 - \frac{1}{2\alpha^1}\Big)C_1\Big)\Big)^2$$

if the parameters α verify the following constraints:

$$\frac{2\alpha^1 \alpha_{10}^1 - \alpha_{20}^1}{4(\alpha^1)^2 - 1} + \frac{2\alpha^2 \alpha_{10}^2 - \alpha_{20}^2}{4(\alpha^2)^2 - 1} \geqslant C_1$$

$$\frac{2\alpha^1\alpha^1_{20} - \alpha^1_{10}}{4(\alpha^1)^2 - 1} + \frac{2\alpha^2\alpha^2_{20} - \alpha^2_{10}}{4(\alpha^2)^2 - 1} \leqslant C_2$$

An expression symmetric to the one above holds for parameters α verifying the following constraints:

$$\frac{2\alpha^1\alpha^1_{10} - \alpha^1_{20}}{4(\alpha^1)^2 - 1} + \frac{2\alpha^2\alpha^2_{10} - \alpha^2_{20}}{4(\alpha^2)^2 - 1} \leqslant C_1$$

$$\frac{2\alpha^1\alpha^1_{20} - \alpha^1_{10}}{4(\alpha^1)^2 - 1} + \frac{2\alpha^2\alpha^2_{20} - \alpha^2_{10}}{4(\alpha^2)^2 - 1} \geqslant C_2$$

Therefore, in the 2-firm-2 period case, the problem consisting of solving Steps 1 and 2 is equivalent to a single level optimization problem, with objective function given in Proposition 6.6.

Note that this problem is nonlinear, non convex.

7. Conclusion

In this paper, we introduced a transient model for joint dynamic pricing and demand learning. We studied a market in a competitive environment and a capacitated setting, where each firm seeks to learn at each time period the price-demand relationship for itself and its competitors, and set its prices for future periods optimally, while guessing its competitors' future optimal policy.

We first formulated the model as a three-step optimization problem: In Step 1, each firm learns the market equilibrium demands. In Step 2, it finds the most accurate estimate for the price-demand parameters, given the optimal demands determined in Step 1. In Step 3, it sets its own prices and guesses its competitors' revenue-maximizing prices for future periods.

We then showed that Steps 1 and 2 could be reformulated as a Mathematical Program with Equilibrium Constraints: the lower level variational inequality is equivalent to the N best-response problems that were solved in Step 1, whereas the upper level objective function is the sum of the objective functions of the N estimations performed in Step 2.

We then focused on the case of a market with 2 firms, with 2 time periods; we derived closed form solutions for the Nash equilibrium, and showed that the MPEC is therefore equivalent to a single-level optimization problem.

Finally, we gained insights into the behavior of the market in the 2-firm-2period setting and reported on early computational results concerning the approach.

Appendix

A. Proofs of the Propositions of Section 6

A.1 Proof of Proposition 6.2

1. If $\sum_{t=1}^{T} \frac{\alpha_{10}^t - \alpha_{12}^t d_2^t}{2\alpha_{11}^t} < C_1$, then the best-response problem of firm 1 is equivalent to an unconstrained problem. As a consequence, the Karsh-Kuhn-Tucker conditions are necessary and sufficient for optimality, with $\lambda_1 = 0$.

   Therefore, equations (20) and (21) yield: $d_1^t = 0$ or $d_1^t = \frac{\alpha_{10}^t - \alpha_{12}^t d_2^t}{2\alpha_{11}^t}$.

2. Else, if $\sum_{t=1}^{T} \frac{\alpha_{10}^t - \alpha_{12}^t d_2^t}{2\alpha_{11}^t} \geqslant C_1$, then the solution to the unconstrained problem are such that the total demand in the unconstrained case exceed capacity. Therefore, the capacity constraint (22) is tight at optimality: $\sum_{t=1}^{T} d_1^t = C_1$. Hence equation (21) yields: for all t such that $d_1^t > 0$, then:

$$d_1^t = \frac{\alpha_{10}^t - \alpha_{12}^t d_2^t}{2\alpha_{11}^t} - \frac{\lambda_1}{2\alpha_{11}^t}$$

   Then equality in (22) implies

$$\lambda_1 = \left(\sum_{t=1}^{T} \frac{1}{2\alpha_{11}^t} \right)^{-1} \left(\sum_{t=1}^{T} \frac{\alpha_{10}^t - \alpha_{12}^t d_2^t}{2\alpha_{11}^t} - C_1 \right)$$

A.2 Proof of Proposition 6.3

Assume both firms have unused capacity at equilibrium. Then from Proposition 6.2, their best response demand is equal to:

$$d_1^t = \frac{\alpha_{10}^t - \alpha_{12}^t d_2^t}{2\alpha_{11}^t}$$

$$d_2^t = \frac{\alpha_{20}^t - \alpha_{21}^t d_1^t}{2\alpha_{22}^t}$$

As a result, the Nash equilibrium demands solve the following system:

$$\begin{pmatrix} 2\alpha_{11}^t & \alpha_{12}^t \\ \alpha_{21}^t & 2\alpha_{22}^t \end{pmatrix} \begin{pmatrix} d_1^t \\ d_2^t \end{pmatrix} = \begin{pmatrix} \alpha_{10}^t \\ \alpha_{20}^t \end{pmatrix}$$

Let $\Delta^t = 4\alpha^t_{11}\alpha^t_{22} - \alpha^t_{12}\alpha^t_{21}$. Inverting the previous system, then the Nash equilibrium is given by:

$$\begin{pmatrix} d_1^t \\ d_2^t \end{pmatrix} = \frac{1}{\Delta^t}\begin{pmatrix} 2\alpha^t_{22} & -\alpha^t_{12} \\ -\alpha^t_{21} & 2\alpha^t_{11} \end{pmatrix}\begin{pmatrix} \alpha^t_{10} \\ \alpha^t_{20} \end{pmatrix}$$

This holds whenever the total demand is less than the capacity for both firms. In other words, the set of parameters α for which this holds is given by:

$$\sum_{t=1}^{T} \frac{2\alpha^t_{22}\alpha^t_{10} - \alpha^t_{12}\alpha^t_{20}}{4\alpha^t_{11}\alpha^t_{22} - \alpha^t_{21}\alpha^t_{12}} \leqslant C_1$$

$$\sum_{t=1}^{T} \frac{2\alpha^t_{11}\alpha^t_{20} - \alpha^t_{21}\alpha^t_{10}}{4\alpha^t_{11}\alpha^t_{22} - \alpha^t_{21}\alpha^t_{12}} \leqslant C_2$$

For $T = 2$ time periods, the system of equations corresponding to the Nash equilibrium is therefore:

$$\begin{pmatrix} 2\alpha^1_{11} & \alpha^1_{12} & 0 & 0 \\ \alpha^1_{21} & 2\alpha^1_{22} & 0 & 0 \\ 0 & 0 & 2\alpha^2_{11} & \alpha^2_{12} \\ 0 & 0 & \alpha^2_{21} & 2\alpha^2_{22} \end{pmatrix}\begin{pmatrix} d_1^1 \\ d_2^1 \\ d_1^2 \\ d_2^2 \end{pmatrix} = \begin{pmatrix} \alpha^1_{10} \\ \alpha^1_{20} \\ \alpha^2_{10} \\ \alpha^2_{20} \end{pmatrix}$$

Or equivalently:

$$\begin{pmatrix} d_1^1 \\ d_2^1 \\ d_1^2 \\ d_2^2 \end{pmatrix} = \begin{pmatrix} 2\alpha^1_{22}/\Delta^1 & -\alpha^1_{12}/\Delta^1 & 0 & 0 \\ -\alpha^1_{21}/\Delta^1 & 2\alpha^1_{11}/\Delta^1 & 0 & 0 \\ 0 & 0 & 2\alpha^2_{22}/\Delta^2 & -\alpha^2_{12}/\Delta^2 \\ 0 & 0 & -\alpha^2_{21}/\Delta^2 & 2\alpha^2_{11}/\Delta^2 \end{pmatrix}\begin{pmatrix} \alpha^1_{10} \\ \alpha^1_{20} \\ \alpha^2_{10} \\ \alpha^2_{20} \end{pmatrix}$$

where $\Delta^t = 4\alpha^t_{11}\alpha^t_{22} - \alpha^t_{12}\alpha^t_{21}$

A.3 Proof of Proposition 6.4

Firm 1's best-response demands are given by equation (27), and similarly for firm 2. Hence for T time periods, the Nash equilibrium demands solve the following system of equations: $\forall\, t = 1, \ldots, T$:

$$2\alpha^t_{11}\left(\sum_{\tau=1}^{T} 1/(2\alpha^\tau_{11})\right) d_1^t + \alpha^t_{12}\left(\sum_{\tau\neq t} 1/(2\alpha^\tau_{11})\right) d_2^t$$

$$-\sum_{\tau\neq t}\left(\frac{\alpha_{12}^{\tau}}{2\alpha_{11}^{\tau}}\right)d_2^{\tau}=C_1+\sum_{\tau=1}^{T}\frac{\alpha_{10}^{t}-\alpha_{10}^{\tau}}{2\alpha_{11}^{\tau}}$$

$$2\alpha_{22}^{t}\left(\sum_{\tau=1}^{T}1/(2\alpha_{22}^{\tau})\right)d_2^{t}+\alpha_{21}^{t}\left(\sum_{\tau\neq t}1/(2\alpha_{22}^{\tau})\right)d_1^{t}$$

$$-\sum_{\tau\neq t}\left(\frac{\alpha_{21}^{\tau}}{2\alpha_{22}^{\tau}}\right)d_1^{\tau}=C_2+\sum_{\tau=1}^{T}\frac{\alpha_{20}^{t}-\alpha_{20}^{\tau}}{2\alpha_{22}^{\tau}}$$

Note that the equations giving the best-response demand for each time period are coupled to other time periods.

For 2 periods, this yields the following system:

$$\begin{pmatrix} 2(\alpha_{11}^1+\alpha_{11}^2) & \alpha_{12}^1 & 0 & -\alpha_{12}^2 \\ \alpha_{21}^1 & 2(\alpha_{22}^1+\alpha_{22}^2) & -\alpha_{21}^2 & 0 \\ 0 & -\alpha_{12}^1 & 2(\alpha_{11}^1+\alpha_{11}^2) & \alpha_{12}^2 \\ -\alpha_{21}^1 & 0 & \alpha_{21}^2 & 2(\alpha_{22}^1+\alpha_{22}^2) \end{pmatrix}\begin{pmatrix} d_1^1 \\ d_2^1 \\ d_1^2 \\ d_2^2 \end{pmatrix}=$$
$$\begin{pmatrix} 2\alpha_{11}^2C_1+\alpha_{10}^1-\alpha_{10}^2 \\ 2\alpha_{22}^2C_2+\alpha_{20}^1-\alpha_{20}^2 \\ 2\alpha_{11}^1C_1+\alpha_{10}^2-\alpha_{10}^1 \\ 2\alpha_{22}^1C_2+\alpha_{20}^2-\alpha_{20}^1 \end{pmatrix}$$

Let $\bar{\alpha}_{ij}=\alpha_{ij}^1+\alpha_{ij}^2$. The determinant of the matrix is: $\Delta=4\bar{\alpha}_{11}\bar{\alpha}_{22}-\bar{\alpha}_{12}\bar{\alpha}_{21}$. The system can then be written as:

$$\begin{pmatrix} d_1^1 \\ d_2^1 \\ d_1^2 \\ d_2^2 \end{pmatrix}=\frac{1}{\Delta}\begin{pmatrix} 2\bar{\alpha}_{22}-\frac{\alpha_{21}^2\bar{\alpha}_{12}}{2\bar{\alpha}_{11}} & -\alpha_{12}^1 & -\frac{\alpha_{21}^2\bar{\alpha}_{12}}{2\bar{\alpha}_{11}} & \alpha_{12}^2 \\ -\alpha_{21}^1 & 2\bar{\alpha}_{11}-\frac{\alpha_{12}^2\bar{\alpha}_{21}}{2\bar{\alpha}_{22}} & \alpha_{21}^2 & -\frac{\alpha_{12}^2\bar{\alpha}_{21}}{2\bar{\alpha}_{22}} \\ -\frac{\alpha_{21}^1\bar{\alpha}_{12}}{2\bar{\alpha}_{11}} & \alpha_{21}^1 & 2\bar{\alpha}_{22}-\frac{\alpha_{21}^1\bar{\alpha}_{12}}{2\bar{\alpha}_{11}} & -\alpha_{12}^2 \\ \alpha_{21}^1 & -\frac{\alpha_{12}^1\bar{\alpha}_{21}}{2\bar{\alpha}_{22}} & -\alpha_{21}^2 & 2\bar{\alpha}_{11}-\frac{\alpha_{12}^1\bar{\alpha}_{21}}{2\bar{\alpha}_{22}} \end{pmatrix}$$
$$\times\begin{pmatrix} 2\alpha_{11}^2C_1+\alpha_{10}^1-\alpha_{10}^2 \\ 2\alpha_{22}^2C_2+\alpha_{20}^1-\alpha_{20}^2 \\ 2\alpha_{11}^1C_1+\alpha_{10}^2-\alpha_{10}^1 \\ 2\alpha_{22}^1C_2+\alpha_{20}^2-\alpha_{20}^1 \end{pmatrix}$$

In order for this to hold, one must verify that the unconstrained Nash equilibrium yields total demands that exceed the firms'capacities:

$$\sum_{t=1}^{T} \frac{2\alpha_{22}^t \alpha_{10}^t - \alpha_{12}^t \alpha_{20}^t}{4\alpha_{11}^t \alpha_{22}^t - \alpha_{21}^t \alpha_{12}^t} \geqslant C_1$$

$$\sum_{t=1}^{T} \frac{2\alpha_{11}^t \alpha_{20}^t - \alpha_{21}^t \alpha_{10}^t}{4\alpha_{11}^t \alpha_{22}^t - \alpha_{21}^t \alpha_{12}^t} \geqslant C_2$$

A.4 Proof of Proposition 6.5

Without loss of generality, assume firm 1's capacity is tight at equilibrium, and firm 2's is not. Hence firm 1's best response demands are given by equation (27), whereas firm 2's best response demands are given by equation (26) where the indices 1 and 2 have been switched. Therefore, for 2 time periods, the Nash equilibrium demands solve the following system of equations:

$$\left(2\left(\alpha_{11}^1 + \alpha_{11}^2\right) - \frac{\alpha_{12}^1 \alpha_{21}^1}{2\alpha_{22}^1} - \frac{\alpha_{12}^2 \alpha_{21}^2}{2\alpha_{22}^2}\right) d_1^1 = \alpha_{10}^1 - \alpha_{10}^2 + \left(2\alpha_{11}^2 - \frac{\alpha_{12}^2 \alpha_{21}^2}{2\alpha_{22}^2}\right) C_1 - \left(\frac{\alpha_{12}^1 \alpha_{20}^1}{2\alpha_{22}^1} - \frac{\alpha_{12}^2 \alpha_{20}^2}{2\alpha_{22}^2}\right)$$

$$\left(2\left(\alpha_{11}^1 + \alpha_{11}^2\right) - \frac{\alpha_{12}^1 \alpha_{21}^1}{2\alpha_{22}^1} - \frac{\alpha_{12}^2 \alpha_{21}^2}{2\alpha_{22}^2}\right) d_1^2 = \alpha_{10}^2 - \alpha_{10}^1 + \left(2\alpha_{11}^1 - \frac{\alpha_{12}^1 \alpha_{21}^1}{2\alpha_{22}^1}\right) C_1 - \left(\frac{\alpha_{12}^2 \alpha_{20}^2}{2\alpha_{22}^2} - \frac{\alpha_{12}^1 \alpha_{20}^1}{2\alpha_{22}^1}\right)$$

Denote by $\bar{d_1}^t$, $\bar{d_s 2}^t$ the solution to the market equilibrium in the unconstrained case. Then inverting the system yields the following expressions for the Nash equilibrium demands:

$$d_1^1 = \left(\frac{\Delta^1}{2\alpha_{22}^1} + \frac{\Delta^2}{2\alpha_{22}^2}\right)^{-1} \left(\frac{\Delta^1}{2\alpha_{22}^1}\bar{d_1}^1 - \frac{\Delta^2}{2\alpha_{22}^2}\bar{d_2}^2 + \frac{\Delta^2}{2\alpha_{22}^2}C_1\right)$$

$$d_1^2 = \left(\frac{\Delta^1}{2\alpha_{22}^1} + \frac{\Delta^2}{2\alpha_{22}^2}\right)^{-1} \left(\frac{\Delta^2}{2\alpha_{22}^2}\bar{d_1}^2 - \frac{\Delta^1}{2\alpha_{22}^1}\bar{d_2}^1 + \frac{\Delta^1}{2\alpha_{22}^1}C_1\right)$$

$$d_2^1 = \frac{\alpha_{20}^1}{2\alpha_{11}^1} - \frac{\alpha_{21}^1}{2\alpha_{22}^1}\left(\frac{\Delta^1}{2\alpha_{22}^1} + \frac{\Delta^2}{2\alpha_{22}^2}\right)^{-1}\left(\frac{\Delta^1}{2\alpha_{22}^1}\bar{d}_1^{\,1} - \frac{\Delta^2}{2\alpha_{22}^2}\bar{d}_2^{\,2} + \frac{\Delta^2}{2\alpha_{22}^2}C_1\right)$$

$$d_2^2 = \frac{\alpha_{20}^2}{2\alpha_{11}^2} - \frac{\alpha_{21}^2}{2\alpha_{22}^2}\left(\frac{\Delta^1}{2\alpha_{22}^1} + \frac{\Delta^2}{2\alpha_{22}^2}\right)^{-1}\left(\frac{\Delta^2}{2\alpha_{22}^2}\bar{d}_1^{\,2} - \frac{\Delta^1}{2\alpha_{22}^1}\bar{d}_2^{\,1} + \frac{\Delta^1}{2\alpha_{22}^1}C_1\right)$$

Furthermore, for this to hold, it must be that the parameters α lie in the following set:

$$\sum_{t=1}^{T} \frac{2\alpha_{22}^t\alpha_{10}^t - \alpha_{12}^t\alpha_{20}^t}{4\alpha_{11}^t\alpha_{22}^t - \alpha_{21}^t\alpha_{12}^t} \geqslant C_1$$

$$\sum_{t=1}^{T} \frac{2\alpha_{11}^t\alpha_{20}^t - \alpha_{21}^t\alpha_{10}^t}{4\alpha_{11}^t\alpha_{22}^t - \alpha_{21}^t\alpha_{12}^t} \leqslant C_2$$

References

Achabal, D. D. and Smith, S. A. Clearance Pricing and Inventory Policies for Retail Chains. *Management Science*, 44(3): 285–300, 1998.

Adida, E. and Perakis, G. A Nonlinear Fluid Model of Dynamic Pricing and Inventory Control with no Backorders. To Appear in *Operations Research*, 2005.

Aviv, Y. and Pazgal, A. *A Partially observed Markov decision process for dynamic pricing*. April, 2004.

Aviv, Y. and Pazgal, A. *Pricing of short life-cycle products through active learning*. April, 2004.

Balvers, R. J. and Cosimano, T. F. Actively learning about Demand and the Dynamics of Price Adjustment. *Econom. Journal*, 100: 882–898, 1990.

Barto, A. and Sutton, R. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 1998.

Bernstein, F. and Federgruen, A. Pricing and Replenishment Strategies in a Distribution System with Competing Retailers. *Working Paper*, 1999.

Bertsimas, D. J. and Perakis, G. Dynamic Pricing: A Learning Approach. In Lawphongpanish, S., Hearn, D. W., and Smith, M. J. (Eds.), *Mathematical and Regional Models for Congestion Charging*, Vol. 101 of *Applied Optimization*. Springer, Regionalwissenschaften, 2006.

Bitran, G. and Caldentey, R. An Overview of Pricing Models for Revenue Management. *Manufacturing and Service Operations Management*, (to appear), 2002.

Bitran, G., Caldentey, R. and Mondschein, S. Coordinating Clearance Markdown sale of Seasonal Products in Retail Chains. *Operations Research*, 46: 609–624, 1998.

Bitran, G. and Mondschein, S. Periodic Pricing of Seasonal Products in Retailing. *Management Science*, 43, (1): 64–79, 1997.

Boyd, S. and Lobo, M. Pricing and Learning under Uncertain Demand. *Working Paper*, Fuqua School of Business, Duke University, 2003.

Carvalho, A. and Puterman, M. How Should a Manager Set Prices When the Demand Function is Unknown. *Technical Report*, Sauder School of Business, University of British Columbia, Vancouver, Canada, 2004.

Dada, M. and Petruzzi, N. C. Dynamic Pricing and Inventory control with Learning. *Naval Research Logistics*, 49: 303–325, 2002.

Dafermos, S. Traffic Equilibrium and Variational Inequalities. *Transportation Science*, 14: 42–54, 1980.

Dockner, E. and Jørgensen, S. Optimal Pricing Strategies for New Products in Dynamic Oligopolies. *Marketing Science*, 7 (4): 315–334, 1988.

Easley, D. and Kiefer, N. Controlling a stochastic process with unknown parameters. *Econometrica*, 56: 1045–64, 1988.

Elmaghraby, W., Guleu, A. and Keskinocak, P. Optimal Markdown Mechanisms in the Presence of Rational Customers with Multi-unit Demand. *Working Paper*, Georgia Institute of Technology, 2002.

Elmaghraby, W. and Keskinocak, P. Dynamic Pricing in the Presence of Inventory Considerations: Research Overview, Current Practices and Future Directions. *Management Science*, 49 (10): 1287–1309, 2003.

Federgruen, A. and Heching, A. Combined Pricing and Inventory Control under Uncertainty. *Operations Research*, 47 (3): 454–475, 1999.

Feng, Y. and Gallego, G. Optimal Starting Times for End-of-season Sales and Optimal Stopping Times for Promotional Fares. *Management Science*, 41: 1371–1391, 1995.

Feng, Y. and Xiao, B. Optimal Policies of Yield Management with Multipled Predetermined Prices. *Management Science*, 48: 332–343, 2000.

Friedman, J. *Oligopoly Theory*. Cambridge University Press, 1983.

Fudenberg, D. and Tirole, J. *Dynamic models of oligopoly*. Harwood Academic Publishers, New York, 1986.

Gallego, G. and van Ryzin, G. A Multiproduct Dynamic Pricing Problem and Its Applications to Network Yield Management. *Operations Research*, 45 (1): 24–41, 1997.

Kachani, S. and Perakis, G. A Fluid Dynamics Model of Dynamic Pricing and Inventory Control for Make-to-Stock Manufacturing Systems. *Working Paper*, Operations Research Center, MIT, 2002.

Kalyanam, K. Pricing Decision under Uncertainty: A Bayesian Mixure Model Approach. *Marketing Science*, 15 (3): 207–221, 1996.

Lazear, E. P. Retail Pricing and Clearance Sales. *The American Economic Review*, 76 (1): 14–32, 1986.

Maglaras, C. and Meissner, J. Dynamic Pricing Strategies for Multiproduct Revenue Management Problems. *Working Paper*, Columbia University, New York, NY, 2004.

Maskin, E. and Tirole, J. A Theory of Dynamic Oligopoly II: Price Competition, kinked demand curves, Edgeworth cycles. *Econometrica*, 56: 571–99, 1988.

McGill, J. I. and van Ryzin, G. Revenue Management: Research Overview and Prospects. *Transportation Science: Focused Issue on Yield Management in Transportation*, 33 (2), 1999.

Mirman, L., Samuelson, L. and Urbano, A. Monopoly Experimentation. *International Economic Review*, 34 (3): 549–63, August 1995.

Perakis, G. and Sood, A. Competitive multi-period pricing for perishable products; a robust optimization approach. To appear in *Mathematical Programming*, 2005.

Popescu, Y. and Wu, Y. Dynamic Pricing Strategies Under Repeated Interactions. *Working Paper*, INSEAD, Fontainebleau, France, 2005.

Rajan, A., Rakash and Steinberg, R. Dynamic Pricing and Ordering Decisions by a Monopolist. *Management Science*, 48: 240–262, 1992.

Rice, J. *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, California, 1995.

Rustichini, A. and Wolinsky, A. Learning About Variable Demand in the Long Run. *Journal of Economic Dynamics and Control*, 19: 1283–92, 1995.

Stigler, G. J. The Kinky Oligopoly Demand Curve and Rigid Prices. *The Journal of Political Economy*, 55 (5): 432–449, 1947.

Sweezy, P. M. Demand Under Conditions of Oligopoly. *The Journal of Political Economy*, 47 (4): 568–573, 1939.

Talluri, K. and van Ryzin, G. *The Theory and Practice of Revenue Management*. Kluwers Academic Publishers, 2004.

Thowsen, G. A dynamic nonstationary inventory problem for a price/quantity setting firm. *Naval Research Logistics Quaterly*, 47:461–476, 1975.

Vives, X. *Oligopoly Pricing - Old Ideas and New Tools*. MIT Press, Cambridge, 1999.

Wardrop, J. G. Some Theoretical Aspects of Road Traffic Research. In *Proceedings of the Institute of Civil Engineers, Part II*, pages 325–378, 1952.

Weatherford, L. R. and Bodily, S. E. A Taxonomy and Research Overview of Perishable-Asset Revenue Management: Yield Management, Overbooking and Pricing. *Operations Research*, 40: 831–844, 1992.

Zabel, E. Monopoly and Uncertainty. *The Review of Economic Studies*, 37 (2): 205–219, 1970.

Chapter 12

An Evolutionary Variational Inequality Formulation of Supply Chain Networks with Time-Varying Demands

Anna Nagurney[1] and Zugang Liu[2]

[1]Radcliffe Institute Fellow, Radcliffe Institute for Advanced Study, 34 Concord Avenue, Harvard University, Cambridge, Massachusetts 02138

[1,2]Department of Finance and Operations Management, Isenberg School of Management, University of Massachusetts, Amherst, Massachusetts 01003

Abstract This paper first develops a multitiered supply chain network equilibrium model with fixed demands and proves that the governing equilibrium conditions satisfy a finite-dimensional variational inequality. The paper then establishes that the static supply chain network model with its governing equilibrium conditions can be reformulated as a transportation network equilibrium model over an appropriately constructed abstract network or supernetwork. This identification provides a new interpretation of equilibrium in supply chain networks with fixed demands in terms of path flows. The equivalence is then further exploited to construct a dynamic supply chain network model with time-varying demands (and flows) using an evolutionary (time-dependent) variational inequality formulation. Recent theoretical results in the unification of projected dynamical systems and evolutionary variational inequalities are presented and then applied to formulate dynamic numerical supply chain network examples and to compute the curves of equilibria. An example with step-wise time-dependent demand is also given for illustration purposes.

Keywords: supply chain; variational inequalities; evolutionary processes; dynamics

1. Introduction

Transportation science has pushed the frontiers in the development and application of rigorous methodologies for the modeling, analysis, and solution of complex network-based problems in which humans interact with critical infrastructure as well as with available technologies. Seminal contributions have been made by numerous authors. Beckmann, McGuire, and Winsten (1956) first rigorously mathematically formulated the traffic assignment or transportation network equilibrium problem and showed that the equilibrium conditions in which no user has any incentive to alter his route of travel coincided with the Kuhn-Tucker conditions of an appropriately constructed optimization problem. Dafermos and Sparrow (1969), subsequently, coined the terms "user-optimization" and "system-optimization" to distinguish between solutions corresponding to, respectively, Wardrop's (1952) first and second principles of travel behavior and also provided algorithms that exploited the network structure. Dafermos (1980) identified the Smith (1979) formulation of transportation network equilibrium as a finite-dimensional variational inequality, an identification that revolutionized the modeling of a spectrum of network equilibrium problems and applications in different disciplines. For an overview of the impact of the Beckmann, McGuire, and Winsten (1956) book, see the paper by Boyce, Mahmassani, and Nagurney (2005). For an overview of finite-dimensional variational inequalities and network-based applications, see the book by Nagurney (1999) and the references therein.

Transportation science has also spearheaded the development of mathematical frameworks to capture disequilibrium behavior associated with the dynamics of users engaged in selecting their routes of travel between origins and destinations. For example, projected dynamical systems theory (cf. Dupuis and Nagurney (1993) and Nagurney and Zhang (1996) and the references therein) was developed, in part, in order to model the behavior of travelers prior to the achievement of an equilibrium state as formulated by a variational inequality problem. The book by Ran and Boyce (1996) contains an overview of dynamic transportation network models and algorithms, along with references to the literature to that date. Formulations of dynamic models of multilayer networks can be found in Nagurney and Dong (2002) and in Zhang, Peeta, and Friesz (2005).

Recently, Cojocaru, Daniele, and Nagurney (2005) built the basis for the merging of projected dynamical systems and that of evolutionary (time-dependent) variational inequalities (which are infinite-dimensional). These two theories had been developed and advanced in parallel, in order to further develop the theoretical analysis and computation of solutions to problems, often, network-based, in which dynamics plays a central role. Cojocaru, Daniele, and Nagurney (2006) demonstrated that the merger of these two theories allows for the modeling of problems that present two theoretically distinct timeframes. They provided the formulation of the associated double-dynamics theory, discussed the question of uniqueness of such curves of equilibria, and also provided conditions for stability properties of such curves in a given neighborhood.

In this paper, we focus on the interplay of transportation network equilibrium models and supply chain network problems and we demonstrate how theory derived from the former class of problems can be used to provide entirely new interpretations of the latter as well as to inform a new kind of time-dependent modeling framework, which is motivated by the unification of projected dynamical systems theory and evolutionary variational inequalities. This paper is organized as follows. In Section 2, we present the multitiered supply chain network equilibrium model with fixed demands, which is motivated by the supply chain network equilibrium model proposed by Nagurney, Dong, and Zhang (2002). In Section 3, we then recall the fixed demand transportation network equilibrium model of Dafermos (1980) and Smith (1979), along with the path-based and link-based variational inequality formulations, which are finite-dimensional. In Section 4, we establish that the supply chain network equilibrium model of Section 2 can be reformulated as a transportation network equilibrium model as described in Section 3, over an appropriately constructed abstract network or *supernetwork* (cf. Nagurney and Dong (2002) and the references therein). A similar equivalence was made by Nagurney (2006) for the case of elastic demands.

In Section 4, we utilize the recently developed unification of the theories of projected dynamical systems and evolutionary variational inequalities to construct a dynamic supply chain network model with time-dependent demands, which is viewed as a dynamic transportation network problem. We also provide some theoretical results. In Section 5, we discuss the computation of curves of equilibria and we illustrate the modeling framework through several numerical dynamic supply chain numerical examples, including one with a step-wise demand function.

2. The Supply Chain Network Model with Fixed Demands

In this Section, we develop a fixed demand version of the supply chain network model proposed in Nagurney, Dong, and Zhang (2002). The model consists of m manufacturers, n retailers, and o demand markets, as depicted in Figure 1. We denote a typical manufacturer by i, a typical retailer by j, and a typical demand market by k. The links in the supply chain network represent transportation/transaction links. The majority of the needed notation is shown in Table 1. The equilibrium solution is denoted by "*". All vectors are assumed to be column vectors, except where noted.

The top-tiered nodes in Figure 1 represent the manufacturers, who produce a homogeneous product and sell to the retailers in the second tier. The consumers at the demand markets are represented by the nodes in the bottom tier of the supply chain network and they acquire the product from the retailers.

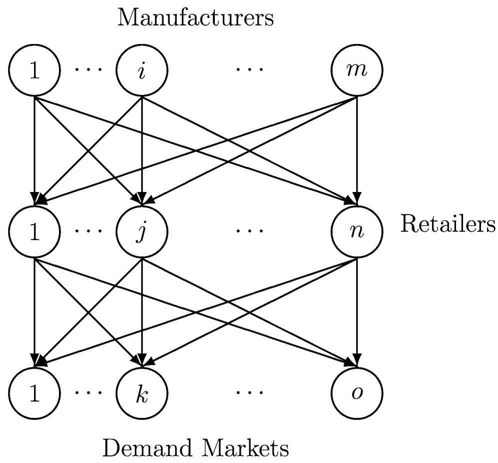


Figure 1. The Network Structure of the Supply Chain.

Table 1. Notation for the Supply Chain Network Equilibrium Model.

Notation	Definition
q	vector of the manufacturers' production outputs with components: $q_1, \ldots, q_m$
Q^1	mn-dimensional vector of product flows transacted/shipped between manufacturers and retailers with component ij: q_{ij}
Q^2	no-dimensional vector of product flows transacted/shipped between retailers and the demand markets with component jk: q_{jk}
γ	n-dimensional vector of shadow prices associated with the retailers with component j: γ_j
d	o-dimensional vector of market demand with component k: d_k
$f_i(q) \equiv f_i(Q^1)$	production cost of manufacturer i with marginal production cost with respect to q_i: $\frac{\partial f_i}{\partial q_i}$ and the marginal production cost with respect to q_{ij}: $\frac{\partial f_i(Q^1)}{\partial q_{ij}}$
$c_{ij}(q_{ij})$	transaction cost between manufacturer i and retailer j with marginal transaction cost: $\frac{\partial c_{ij}(q_{ij})}{\partial q_{ij}}$
s	vector of the retailers' supplies of the product with components: $s_1, \ldots, s_n$
$c_j(s) \equiv c_j(Q^1)$	handling cost of retailer j with marginal handling cost with respect to s_j: $\frac{\partial c_j}{\partial s_j}$ and with the marginal handling cost with respect to q_{ij}: $\frac{\partial c_j(Q^1)}{\partial q_{ij}}$
$c_{jk}(Q^2)$	unit transaction cost between retailer j and demand market k

We first describe the behavior of the manufacturers and the retailers. We then discuss the behavior of the consumers at the demand markets. Finally, we state the equilibrium conditions for the supply chain network and provide the finite-dimensional variational inequality governing the equilibrium.

2.1 The Behavior of the Manufacturers and their Optimality Conditions

Let ρ^*_{1ij} denote the price charged for the product by manufacturer i in transacting with retailer j. The price ρ^*_{1ij} is an endogenous variable and will be determined once the entire supply chain network equilibrium model is solved. We assume that the quantity produced by manufacturer i must satisfy the following conservation of flow equation:

$$q_i = \sum_{j=1}^{n} q_{ij}, \tag{1}$$

which states that the quantity of the product produced by manufacturer i is exactly equal to the sum of the quantities transacted/shipped between a manufacturer and the retailers. The production cost function f_i for each manufacturer i; $i = 1, \ldots, m$, as delineated in Table 1, can, in view of (1), be reexpressed as a function of the flows Q^1.

Hence, assuming that the manufacturers are profit-maximizers, we can express the optimization problem faced by manufacturer i as:

$$\text{Maximize} \quad \sum_{j=1}^{n} \rho^*_{1ij} q_{ij} - f_i(Q^1) - \sum_{j=1}^{n} c_{ij}(q_{ij}), \tag{2}$$

subject to: $q_{ij} \geq 0$, for all j; $j = 1, \ldots, n$.

The first term in (2) represents the revenue and the subsequent two terms the production cost and the transaction costs, respectively, for manufacturer i.

We assume that the manufacturers compete in a noncooperative manner in the sense of Cournot (1838) and Nash (1950, 1951), and the production cost functions and the transaction cost functions for each manufacturer are continuously differentiable and convex. The optimality conditions for all manufacturers i; $i = 1, \ldots, m$, simultaneously, can then be expressed as the following variational inequality (cf. Nagurney, Dong and Zhang (2002), Bazaraa, Sherali, and Shetty (1993), Gabay and Moulin (1980); see also Dafermos and Nagurney (1987) and Nagurney (1999)): determine $Q^{1*} \in R^{mn}_+$ satisfying:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \left[\frac{\partial f_i(Q^{1*})}{\partial q_{ij}} + \frac{\partial c_{ij}(q^*_{ij})}{\partial q_{ij}} - \rho^*_{1ij} \right] \times \left[q_{ij} - q^*_{ij} \right] \geq 0, \quad \forall Q^1 \in R^{mn}_+. \tag{3}$$

2.2 The Behavior of the Retailers and their Optimality Conditions

The retailers, in turn, purchase the product from the manufacturers and transact with the consumers at the demand markets. Thus, a retailer is involved in transactions both with the manufacturers as well as with the demand markets.

Let ρ^*_{2j} denote the price charged by the retailer j for the product. This price will be determined endogenously after the the model is solved. We assume that the retailers are also profit-maximizers and, hence, the optimization problem faced by a retailer j is given by:

$$\text{Maximize} \quad \sum_{k=1}^{o} \rho^*_{2j} q_{jk} - c_j(Q^1) - \sum_{i=1}^{m} \rho^*_{1ij} q_{ij} \tag{4}$$

subject to:

$$\sum_{k=1}^{o} q_{jk} \leq \sum_{i=1}^{m} q_{ij}, \tag{5}$$

and the nonnegativity constraints: $q_{ij} \geq 0$, and $q_{jk} \geq 0$, for all i; $i = 1, \ldots, m$, and k; $k = 1, \ldots, o$.

The first term in the objective function (4) represents the revenue whereas the second and third terms represent, respectively, the handling cost and the payout to the manufacturers. Constraint (5) expresses that the total quantity of the product transacted with the demand markets by a retailer cannot exceed the total amount that the retailer has obtained from the manufacturers.

We assume that the retailers also compete in a noncooperative manner and that the handling cost for each retailer is continuously differentiable and convex. Then the optimality conditions for all the retailers simultaneously can be expressed as the variational inequality: determine $(Q^{1*}, Q^{2*}, \gamma^*) \in R_+^{mn+no+n}$ satisfying:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \left[\frac{\partial c_j(Q^{1*})}{\partial q_{ij}} + \rho^*_{1ij} - \gamma^*_j \right] \times \left[q_{ij} - q^*_{ij} \right]$$

$$+ \sum_{j=1}^{n} \sum_{k=1}^{o} \left[-\rho^*_{2j} + \gamma^*_j \right] \times \left[q_{jk} - q^*_{jk} \right]$$

$$+ \sum_{j=1}^{n} \left[\sum_{i=1}^{m} q^*_{ij} - \sum_{k=1}^{o} q^*_{jk} \right] \times \left[\gamma_j - \gamma^*_j \right] \geq 0,$$

$$\forall (Q^1, Q^2, \gamma) \in R_+^{mn+no+n}. \tag{6}$$

As noted in Table 1, the term γ_j is the Lagrange multiplier/shadow price associated with constraint (5) for retailer j and γ is the n-dimensional vector of all the shadow prices.

2.3 The Consumers at the Demand Markets and the Equilibrium Conditions

We now discuss the behavior of the consumers at the demand markets. The consumers take into account the prices charged by the retailers and the unit transaction costs incurred to obtain the product in making their consumption decisions. In the static model, we assume that the demand for the product at each demand market is fixed and known (later in this paper, we will develop the dynamic model in which we allow the demand to be time-varying). Hence, the following conservation of flow equations must hold:

$$d_k = \sum_{j=1}^{n} q_{jk}, \quad k = 1, \ldots, o, \tag{7}$$

where d_k is fixed for each demand market k.

We assume that the unit transaction cost functions c_{jk} are continuous functions for j; $j = 1, \ldots, n$ and k; $k = 1, \ldots, o$, and are of the form given in Table 1.

The equilibrium conditions for consumers at demand market k then take the form: for each retailer j; $j = 1, \ldots, n$:

$$\rho^*_{2j} + c_{jk}(Q^{2*}) \begin{cases} = \rho^*_{3k}, \text{ if } & q^*_{jk} > 0, \\ \geqslant \rho^*_{3k}, \text{ if } & q^*_{jk} = 0. \end{cases} \tag{8}$$

Conditions (8) state that, in equilibrium, if the consumers at demand market k purchase the product from retailer j, then the price the consumers pay is exactly equal to the price charged by the retailer plus the unit transaction cost. However, if the sum of the price charged by the retailer and the unit transaction cost exceeds the price that the consumers are willing to pay at the demand market, there will be no transaction between this retailer/demand market pair. In equilibrium, condition (8) must hold simultaneously for all demand markets. We can express these equilibrium conditions as the following variational inequality: determine $Q^{2*} \in \mathcal{K}^1$, such that

$$\sum_{j=1}^{n} \sum_{k=1}^{o} \left[\rho^*_{2j} + c_{jk}(Q^{2*})\right] \times [q_{jk} - q^*_{jk}] \geq 0, \quad \forall Q^2 \in \mathcal{K}^1, \tag{9}$$

where $\mathcal{K}^1 \equiv \{Q^2 | Q^2 \in R^{no}_+ \text{ and (7) holds}\}$.

In Nagurney, Dong, and Zhang (2002), it was assumed that the demand functions associated with the demand markets were elastic and depended upon the prices of the product at the demand markets. Nagurney (2006), subsequently, proved that the elastic demand supply chain network equilibrium model could be reformulated as an elastic demand transportation network equilibrium model (cf. Dafermos and Nagurney (1984); see also Fisk and Boyce (1983)) over an appropriately constructed supernetwork. That paper, however, did not introduce any dynamics.

2.4 The Equilibrium Conditions of the Supply Chain

In equilibrium, the optimality conditions of all the manufacturers, the optimality conditions of all the retailers, and the equilibrium conditions for all the demand markets must be simultaneously satisfied so that no decision-maker has any incentive to alter his transactions. We now formally state the equilibrium conditions for the entire supply chain network as follows.

DEFINITION 1 (Supply Chain Network Equilibrium (Fixed Demands)). *The equilibrium state of the supply chain network with fixed demands is one where the flows of the product between the tiers of the decision-makers coincide and the flows and prices satisfy the sum of conditions (3), (6), and (9).*

We now state and prove:

THEOREM 1 (Variational Inequality Formulation of the Supply Chain Network Equilibrium). *The equilibrium conditions governing the supply chain network according to Definition 1 coincide with the solution of the (finite-dimensional) variational inequality given by: determine* $(Q^{1*}, Q^{2*}, \gamma^*) \in \mathcal{K}^2$ *satisfying:*

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial f_i(Q^{1*})}{\partial q_{ij}}+\frac{\partial c_{ij}(q^*_{ij})}{\partial q_{ij}}+\frac{\partial c_j(Q^{1*})}{\partial q_{ij}}-\gamma^*_j\right]\times\left[q_{ij}-q^*_{ij}\right]$$
$$+\sum_{j=1}^{n}\sum_{k=1}^{o}\left[c_{jk}(Q^{2*})+\gamma^*_j\right]\times\left[q_{jk}-q^*_{jk}\right]$$
$$+\sum_{j=1}^{n}\left[\sum_{i=1}^{m}q^*_{ij}-\sum_{k=1}^{o}q^*_{jk}\right]\times\left[\gamma_j-\gamma^*_j\right]\geq 0,$$
$$\forall (Q^1, Q^2, \gamma) \in \mathcal{K}^2, \qquad (10)$$

where $\mathcal{K}^2 \equiv \{(Q^1, Q^2, \gamma) | (Q^1, Q^2, \gamma) \in R_+^{mn+no+n}$ *and* (7) *holds*$\}$.

PROOF. We first demonstrate that an equilibrium pattern according to Definition 1 satisfies the variational inequality (10). We sum up inequalities (3), (6), and (9), and, after algebraic simplifications, obtain (10).

We now prove the converse, that is, a solution to variational inequality (10) satisfies the sum of conditions (3), (6), and (9), and is, therefore, a supply chain network equilibrium pattern according to Definition 1.

First, we add the term $\rho^*_{1ij} - \rho^*_{1ij}$ to the first term in the first summand expression in (10). Then, we add the term $\rho^*_{2j} - \rho^*_{2j}$ to the first term in the second summand expression in (10). Because these terms are all equal to zero, they do not change (10) and we obtain the following inequality:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial f_i(Q^{1*})}{\partial q_{ij}} + \frac{\partial c_{ij}(q^*_{ij})}{\partial q_{ij}} + \frac{\partial c_j(Q^{1*})}{\partial q_{ij}} - \gamma^*_j + \rho^*_{1ij} - \rho^*_{1ij}\right] \times \left[q_{ij} - q^*_{ij}\right] + \sum_{j=1}^{n}\sum_{k=1}^{o}\left[c_{jk}(Q^{2*}) + \gamma^*_j + \rho^*_{2j} - \rho^*_{2j}\right] \times \left[q_{jk} - q^*_{jk}\right] + \sum_{j=1}^{n}\left[\sum_{i=1}^{m} q^*_{ij} - \sum_{k=1}^{o} q^*_{jk}\right] \times \left[\gamma_j - \gamma^*_j\right] \geq 0, \quad \forall (Q^1, Q^2, \gamma) \in \mathcal{K}^2, \tag{11}$$

which can be rewritten as:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial f_i(Q^{1*})}{\partial q_{ij}} + \frac{\partial c_{ij}(q^*_{ij})}{\partial q_{ij}} - \rho^*_{1ij}\right] \times \left[q_{ij} - q^*_{ij}\right] + \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial c_j(Q^{1*})}{\partial q_{ij}} + \rho^*_{1ij} - \gamma^*_j\right] \times \left[q_{ij} - q^*_{ij}\right] + \sum_{j=1}^{n}\sum_{k=1}^{o}\left[-\rho^*_{2j} + \gamma^*_j\right] \times \left[q_{jk} - q^*_{jk}\right] + \sum_{j=1}^{n}\left[\sum_{i=1}^{m} q^*_{ij} - \sum_{k=1}^{o} q^*_{jk}\right] \times \left[\gamma_j - \gamma^*_j\right] + \sum_{j=1}^{n}\sum_{k=1}^{o}\left[\rho^*_{2j} + c_{jk}(Q^{2*})\right] \times [q_{jk} - q^*_{jk}] \geqslant 0, \quad \forall (Q^1, Q^2, \gamma) \in \mathcal{K}^2. \tag{12}$$

Clearly, (12) is equal to the sum of the optimality conditions (3) and (6), and the equilibrium conditions (9) and is, hence, a variational inequality governing the supply chain network equilibrium according to Definition 1. □

The variational inequality (10) is different from the variational inequality formulation of elastic demand supply chain network equilibrium problems derived by Nagurney, Dong, and Zhang (2002), as expected, since the feasible set is different and we do not have demand functions, but, rather, we now assume fixed demands.

At the end of Section 4, we describe how to recover the nodal prices in the supply chain network with fixed demands consisting of the top tier prices: ρ^*_{1ij}; for $i=1,\dots,m$; $j=1,\dots,n$; the middle tier prices: ρ^*_{2j}; $j=1,\dots,n$, and the demand market prices: ρ^*_{3k}; $k=1,\dots,o$. The nodal prices of the supply chain network guarantee that the optimality conditions (3), (6), and the equilibrium conditions (8) hold separately at the solution of the variational inequality (10).

The following corollary establishes that, in equilibrium, the supply chain structure is as in Figure 1. Of course, links that have equilibrium flows of zero can, in effect, be eliminated from the supply chain network. This corollary is also useful in establishing the equivalence of the supply chain network equilibrium problem with fixed demands with the transportation network equilibrium problem with fixed demands over an appropriately constructed abstract network or supernetwork, as we demonstrate in the next Section.

COROLLARY 1. *The market for the product clears for each retailer in the supply chain network equilibrium, that is,* $\sum_{i=1}^m q^*_{ij} = \sum_{k=1}^o q^*_{jk}$ *for* $j = 1,\dots,n$.

PROOF. Clearly from (10), we know that, if $\gamma^*_j > 0$, then $\sum_{i=1}^m q^*_{ij} = \sum_{k=1}^o q^*_{jk}$ holds. Now we consider the case where $\gamma^*_j = 0$ for some retailer j. Let us examine the first terms in inequality (10). Since we have assumed that the production cost functions, the transaction cost functions, and the handling cost functions are convex, it is not unreasonable to further assume that either the marginal production cost or the marginal transaction cost or the marginal handling cost for each manufacturer/retailer pair is strictly positive at equilibrium. Then we know that $\frac{\partial f_i(Q^{1*})}{\partial q_{ij}} + \frac{\partial c_{ij}(q^*_{ij})}{\partial q_{ij}} + \frac{\partial c_j(Q^{1*})}{\partial q_{ij}} > 0$, which implies that $q^*_{ij} = 0$, for all i, for that j. It follows then from the third term in (10), that $\sum_{k=1}^o q^*_{jk} = 0$. Hence, we have that $\sum_{k=1}^o q^*_{jk} = 0 = \sum_{i=1}^m q^*_{ij}$ for any j such that $\gamma^*_j = 0$. Therefore, we conclude that the market clears for each retailer in the supply chain equilibrium. □

Since we are interested in the determination of the equilibrium flows and prices, we can transform constraint (5) into:

$$\sum_{k=1}^{o} q_{jk} = \sum_{i=1}^{m} q_{ij}, \quad j = 1, \ldots, n. \tag{13}$$

Now we can define the feasible set as $\mathcal{K}^3 \equiv \{(Q^1, Q^1) \in R_+^{mn+no}$, such that (13) holds$\}$.

In addition, for notational convenience, we let

$$s_j \equiv \sum_{k=1}^{o} q_{jk}, \quad j = 1, \ldots, n. \tag{14}$$

The following results then follow immediately:

COROLLARY 2. *A solution $(Q^{1*}, Q^{2*}) \in \mathcal{K}^3$ to the variational inequality problem:*

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial f_i(Q^{1*})}{\partial q_{ij}} + \frac{\partial c_{ij}(q_{ij}^*)}{\partial q_{ij}} + \frac{\partial c_j(Q^{1*})}{\partial q_{ij}}\right] \times \left[q_{ij} - q_{ij}^*\right]$$
$$+ \sum_{j=1}^{n}\sum_{k=1}^{o}\left[c_{jk}(Q^{2*})\right] \times \left[q_{jk} - q_{jk}^*\right] \geq 0, \quad \forall (Q^1, Q^2) \in \mathcal{K}^3; \tag{15}$$

equivalently, a solution $(q^*, Q^{1*}, s^*, Q^{2*}) \in \mathcal{K}^4$ to:

$$\sum_{i=1}^{m}\left[\frac{\partial f_i(q^*)}{\partial q_i}\right] \times \left[q_i - q_i^*\right] + \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial c_{ij}(q_{ij}^*)}{\partial q_{ij}}\right] \times \left[q_{ij} - q_{ij}^*\right]$$
$$+ \sum_{j=1}^{n}\left[\frac{\partial c_j(s^*)}{\partial s_j}\right] \times \left[s_j - s_j^*\right]$$
$$+ \sum_{j=1}^{n}\sum_{k=1}^{o}\left[c_{jk}(Q^{2*})\right] \times \left[q_{jk} - q_{jk}^*\right] \geq 0, \quad \forall (q, Q^1, s, Q^2) \in \mathcal{K}^4, \tag{16}$$

where $\mathcal{K}^4 \equiv \{(q, Q^1, s, Q^2) | (q, Q^1, s, Q^2) \in R_+^{m+mn+n+no}$ and (1), (7), (13), and, (14) hold$\}$, also satisfies variational inequality (10).

PROOF. We prove Corollary 2 by contradiction. In particular, we demonstrate that if (10) is not true, then (15) does not hold. We assume that for some $(Q^1, Q^2) \in \mathcal{K}^3$ with $\gamma \in R^n_+$ that the left-hand side of (10) is less than or equal to zero, which implies that:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial f_i(Q^{1*})}{\partial q_{ij}}+\frac{\partial c_{ij}(q^*_{ij})}{\partial q_{ij}}+\frac{\partial c_j(Q^{1*})}{\partial q_{ij}}\right]\times\left[q_{ij}-q^*_{ij}\right]$$
$$+\sum_{j=1}^{n}\sum_{k=1}^{o}c_{jk}(Q^{2*})\times\left[q_{jk}-q^*_{jk}\right]$$
$$\leq\sum_{i=1}^{m}\sum_{j=1}^{n}\gamma^*_j\left[q_{ij}-q^*_{ij}\right]-\sum_{j=1}^{n}\sum_{k=1}^{o}\gamma^*_j\left[q_{jk}-q^*_{jk}\right]$$
$$-\sum_{j=1}^{n}\left[\sum_{i=1}^{m}q^*_{ij}-\sum_{k=1}^{o}q^*_{jk}\right]\times\left[\gamma_j-\gamma^*_j\right]. \tag{17}$$

But, after algebraic simplification and the use of Corollary 1, the right-hand side of (17) is reduced to zero. Hence, (15) cannot hold, and the conclusion follows.

We can obtain variational inequality (16) from variational inequality (15) through simple algebraic relationships and the use of (1), (13), and (14). □

3. The Transportation Network Equilibrium Model with Fixed Demands

In this Section, we review the transportation network equilibrium model with fixed demands, due to Smith (1979) and Dafermos (1980).

Consider a network $\mathcal{G}$ with the set of links L with K elements, the set of paths P with Q elements, and the set of origin/destination (O/D) pairs W with Z elements. We denote the set of paths connecting O/D pair w by P_w; the links by a, b, etc; the paths by p, q, etc., and the O/D pairs by w_1, w_2, etc.

The flow on path p is denoted by x_p and the flow on link a by f_a. The travel cost experienced by a user on a path p is denoted by C_p and the travel cost incurred on a link a by c_a. We assume that the user link cost functions are continuous. We also denote the travel demand associated with traveling between O/D pair w by d_w and the travel disutility by λ_w, where d_w is assumed to be fixed and known for all w.

Hence, the following conservation of flow equations must hold:

$$d_w = \sum_{p \in P_w} x_p, \quad \forall w, \tag{18}$$

that is, the travel demand associated with an O/D pair must be equal to the sum of the flows on the paths that connect that O/D pair.

The following conservation of flow equations relate the link flows to the path flows:

$$f_a = \sum_{p \in P} x_p \delta_{ap}, \quad \forall a \in L, \tag{19}$$

where $\delta_{ap} = 1$, if path p contains link a, and $\delta_{ap} = 0$, otherwise. Hence, the flow on a link is equal to the sum of the flows of paths that contain that link.

The user cost on a path is equal to the sum of user costs on links the path consists of, which can be represented by the following:

$$C_p = \sum_{a \in L} c_a \delta_{ap}, \quad \forall p \in P, \tag{20}$$

For the sake of generality, we allow the user cost on a link to depend upon the entire vector of link flows, denoted by f, so that

$$c_a = c_a(f), \quad \forall a \in L. \tag{21}$$

As established by Smith (1979) and Dafermos (1980), a path flow pattern $x^* \in \mathcal{K}^5$, where $\mathcal{K}^5 \equiv \{x | x \in R^Q_+ \text{ and (18) holds}\}$ is said to be a transportation network equilibrium (according to Wardrop's first principle; see Wardrop (1952) and Beckmann, McGuire, and Winsten (1956)), if, once established, no user has any incentive to alter his travel decisions. The state can be expressed by the following equilibrium conditions which must hold for every O/D pair $w \in W$ and every path $p \in P_w$:

$$C_p(x^*) - \lambda^*_w \begin{cases} = 0, \text{ if } & x^*_p > 0, \\ \geq 0, \text{ if } & x^*_p = 0. \end{cases} \tag{22}$$

Conditions (22) express that the user costs of all utilized paths joining an O/D pair are equal and minimal. As described in Dafermos (1980) and Smith (1979) the transportation network equilibrium pattern according to conditions

(22) coincides to the following finite-dimensional variational inequality in path flows: determine $x^* \in \mathcal{K}^5$ such that

$$\sum_{w\in W} \sum_{p\in P_w} C_p(x^*) \times \left[x_p - x_p^*\right] \geq 0, \quad \forall x \in \mathcal{K}^5. \tag{23}$$

We now provide the standard variational inequality form of (23). In particular, we define the function that enters the variational inequality $F(x) \equiv C(x)$ and the feasible set $\mathcal{K} \equiv \mathcal{K}^5$. We then seek to determine $x^* \in \mathcal{K}$ such that

$$\langle F(x^*), x - x^* \rangle \geq 0, \quad \forall x \in \mathcal{K},$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in n-dimensional space where n here is equal to the dimension of path flows, that is, Q. In Section 5, we present a dynamic version of the transportation network equilibrium problem formulated as an evolutionary variational inequality and we will use this standard form in connecting the static formulation with the dynamic one.

We now provide the equivalent variational inequality in link flows, which will be utilized in the demonstration of the supernetwork equivalence in Section 4. For additional background, see the book by Nagurney (1999) and the references therein.

THEOREM 2. *A link flow pattern is a transportation network equilibrium if and only if it satisfies the variational inequality problem: determine $f^* \in \mathcal{K}^6$ satisfying*

$$\sum_{a\in L} c_a(f^*) \times (f_a - f_a^*) \geq 0, \quad \forall f \in \mathcal{K}^6, \tag{24}$$

where $\mathcal{K}^6 \equiv \{f \in R_+^K \,|\, \textit{there exists an } x \textit{ satisfying } (18) \textit{ and } (19)\}$.

The continuity of the link cost functions and the compactness of the feasible sets $\mathcal{K}^5$ and $\mathcal{K}^6$ guarantee the existence of solutions to both variational inequalities (23) and (24) from the standard theory of variational inequalities (see Kinderlehrer and Stampacchia (1980)).

4. Transportation Network Equilibrium Reformulation of Supply Chain Network Equilibrium with Fixed Demands

In this Section, we establish the supernetwork equivalence of the fixed demand supply chain network equilibrium with a properly configured transportation network equilibrium model as discussed in Section 3.

We consider a supply chain network as discussed in Section 2 which consists of m manufacturers: $i = 1, \ldots, m$; n retailers: $j = 1, \ldots, n$, and o demand markets: $k = 1, \ldots, o$. The supernetwork $\mathcal{G}_S$ of the isomorphic transportation network equilibrium model is depicted in Figure 2 and is constructed as follows. The supernetwork $\mathcal{G}_S$ consists of the single origin node 0 at the top tier, and o destination nodes at the bottom tier denoted, respectively, by: $z_1, \ldots, z_o$. Thus, there are o O/D pairs in $\mathcal{G}_S$ denoted, respectively, by $w_1 = (0, z_1)$, ..., $w_k = (0, z_k)$,..., $w_o = (0, z_o)$. Node 0 is connected to each second-tiered node x_i, where $i = 1, \ldots, m$. Each second-tiered node x_i, in turn, is connected to each third-tiered node y_j, where $j = 1, \ldots, n$. Each node y_j, in turn, is connected with a corresponding node y'_j in the fourth tier by a single link. Finally, from each fourth-tiered node y'_j there are o links emanating to the bottom-tiered nodes z_k. There are, hence, $1 + m + 2n + o$ nodes in the supernetwork in Figure 2, $K = m + mn + n + no$ links, $Z = o$ O/D pairs, and $Q = mo$ paths.

We now define the links in the supernetwork in Figure 2 and the associated flows. Let a_i denote the link from node 0 to node x_i with associated link flow f_{a_i}, for $i = 1, \ldots, m$. Let a_{ij} denote the link from node x_i to node y_j with associated link flow $f_{a_{ij}}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Also, let $a_{jj'}$ denote the link connecting node y_j with node $y_{j'}$ with associated link flow $f_{a_{jj'}}$ for

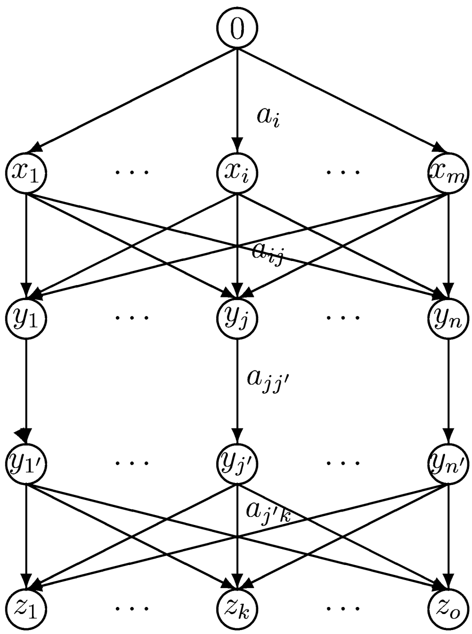


Figure 2. The $\mathcal{G}_S$ Supernetwork Representation of Supply Chain Network Equilibrium with Fixed Demands.

j; $j = 1, \ldots, n$; and j'; $j' = 1, \ldots, n$. Finally, let $a_{j'k}$ denote the link joining node $y_{j'}$ with node z_k for $j' = 1', \ldots, n'$ and $k = 1, \ldots, o$ and with associated link flow $f_{a_{j'k}}$. We group the $\{f_{a_i}\}$ into the vector f^1; the $\{f_{a_{ij}}\}$ into the vector f^2; the $\{f_{a_{jj'}}\}$ into the vector f^3, and the $\{f_{a_{j'k}}\}$ into the vector f^4.

Hence, a typical path in $\mathcal{G}_S$, $p_{ijj'k}$, consists of four links: a_i, a_{ij}, $a_{jj'}$, and $a_{j'k}$. We denote the path flow associated with path $p_{ijj'k}$ by $x_{p_{ijj'k}}$. Also, we let d_{w_k} denote the known fixed demand associated with O/D pair w_k and we let λ_{w_k} denote the travel disutility associated with O/D pair w_k.

We assume that the link flows satisfy the conservation of flow equations (19), that is:

$$f_{a_i} = \sum_{j=1}^{n} \sum_{j'=1'}^{n'} \sum_{k=1}^{o} x_{p_{ijj'k}}, \quad i = 1, \ldots, m, \tag{25}$$

$$f_{a_{ij}} = \sum_{j'=1'}^{n'} \sum_{k=1}^{o} x_{p_{ijj'k}}, \quad i = 1, \ldots, m; \; j = 1, \ldots, n, \tag{26}$$

$$f_{a_{jj'}} = \sum_{i=1}^{m} \sum_{k=1}^{o} x_{p_{ijj'k}}, \quad j = 1, \ldots, n; \; j' = 1, \ldots, n, \tag{27}$$

$$f_{a_{j'k}} = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{p_{ijj'k}}, \quad j' = 1, \ldots, n; \; k = 1, \ldots, o. \tag{28}$$

Also, we have that

$$d_{w_k} = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{j'=1'}^{n'} x_{ijj'k}, \quad k = 1, \ldots, o. \tag{29}$$

A path flow pattern induces a feasible link flow pattern if all path flows are nonnegative and (25)–(29) are satisfied.

Given a feasible product shipment/transaction pattern for the supply chain model with fixed demands, $(q, Q^1, s, Q^2) \in \mathcal{K}^4$, we may construct a feasible link flow pattern on the network $\mathcal{G}_S$ as follows: the link flows are defined as:

$$q_i \equiv f_{a_i}, \quad i = 1, \ldots, m, \tag{29}$$

$$q_{ij} \equiv f_{a_{ij}}, \quad i = 1, \ldots, m; \; j = 1, \ldots, n, \tag{30}$$

$$s_j \equiv f_{a_{jj'}}, \quad j = 1, \ldots, n; \; j' = 1, \ldots, n', \tag{31}$$

$$q_{jk} = f_{a_{j'k}}, \quad j = 1, \ldots, n; \; j' = 1', \ldots, n'; \; k = 1, \ldots, o. \tag{32}$$

Note that if (q, Q^1, s, Q^2) is feasible then the link flow pattern constructed according to (30)–(33) is also feasible and the corresponding path flow pattern that induces such a link flow pattern is, hence, also feasible.

We now assign travel costs on the links of the network $\mathcal{G}_S$ as follows: with each link a_i we assign a travel cost c_{a_i} defined by

$$c_{a_i} \equiv \frac{\partial f_i}{\partial q_i}, \quad i = 1, \ldots, m; \tag{34}$$

with each link a_{ij} we assign a travel cost $c_{a_{ij}}$ defined by:

$$c_{a_{ij}} \equiv \frac{\partial c_{ij}}{\partial q_{ij}}, \quad i = 1, \ldots, m; j = 1, \ldots, n; \tag{35}$$

and with each link jj' we assign a travel cost defined by

$$c_{a_{jj'}} \equiv \frac{\partial c_j}{\partial s_j}, \quad j = 1, \ldots, n; j' = 1, \ldots, n. \tag{36}$$

Finally, for each link $a_{j'k}$ we assign a travel cost defined by

$$c_{a_{j'k}} \equiv c_{jk}, \quad j' = 1, \ldots, n'; k = 1, \ldots, o. \tag{37}$$

Hence, a traveler traveling on path $p_{ijj'k}$ experiences a travel cost $C_{p_{ijj'k}}$ given by

$$C_{p_{ijj'k}} = c_{a_i} + c_{a_{ij}} + c_{a_{jj'}} + c_{a_{j'k}} = \frac{\partial f_i}{\partial q_i} + \frac{\partial c_{ij}}{\partial q_{ij}} + \frac{\partial c_j}{\partial s_j} + c_{jk}. \tag{38}$$

Also, we define the travel demands associated with the O/D pairs as follows:

$$d_{w_k} \equiv d_k, \quad k = 1, \ldots, o \tag{39}$$

and the travel disutilities:

$$\lambda_{w_k} \equiv \rho_{3k}, \quad k = 1, \ldots, o. \tag{40}$$

Consequently, according to the fixed demand transportation network equilibrium conditions (22), we have that, for each O/D pair w_k in $\mathcal{G}_S$ and every path connecting the O/D pair w_k, the following conditions must hold:

$$C_{p_{ijj'k}} - \lambda^*_{w_k} = \frac{\partial f_i}{\partial q_i} + \frac{\partial c_{ij}}{\partial q_{ij}} + \frac{\partial c_j}{\partial q_j} + c_{jk} - \lambda^*_{w_k} \begin{cases} = 0, \text{ if } & x^*_{p_{ijj'k}} > 0, \\ \geq 0, \text{ if } & x^*_{p_{ijj'k}} = 0, \end{cases} \tag{41}$$

where

$$\sum_{p \in P_{w_k}} x^*_{p_{ijj'k}} = d_{w_k}. \tag{42}$$

We now provide the variational inequality formulation of the equilibrium conditions (41) in link form as in (24). A link flow pattern $f^* \in \mathcal{K}^6$ is an equilibrium according to (41), if and only if it satisfies:

$$\sum_{i=1}^{m} c_{a_i}(f^{1*}) \times (f_{a_i} - f^*_{a_i}) + \sum_{i=1}^{m}\sum_{j=1}^{n} c_{a_{ij}}(f^{2*}) \times (f_{a_{ij}} - f^*_{a_{ij}})$$

$$+ \sum_{j=1}^{n}\sum_{j'=1'}^{n'} c_{a_{jj'}}(f^{3*}) \times (f_{a_{jj'}} - f^*_{a_{jj'}})$$

$$+ \sum_{j'=1}^{n'}\sum_{k=1}^{n} c_{a_{j'k}}(f^{4*}) \times (f_{a_{j'k}} - f^*_{a_{j'k}}) \geq 0, \quad \forall f \in \mathcal{K}^6, \tag{43}$$

which, through expressions (30)–(33), and (34)–(37) yields:

$$\sum_{i=1}^{m} \left[\frac{\partial f_i(q^*)}{\partial q_i}\right] \times \left[q_i - q_i^*\right] + \sum_{i=1}^{m}\sum_{j=1}^{n} \left[\frac{\partial c_{ij}(q^*_{ij})}{\partial q_{ij}}\right] \times \left[q_{ij} - q^*_{ij}\right]$$

$$+ \sum_{j=1}^{n} \left[\frac{\partial c_j(s^*)}{\partial s_j}\right] \times \left[s_j - s_j^*\right]$$

$$+ \sum_{j=1}^{n}\sum_{k=1}^{o} \left[c_{jk}(Q^{2*})\right] \times \left[q_{jk} - q^*_{jk}\right] \geq 0,$$

$$\forall (q, Q^1, s, Q^2) \in \mathcal{K}^4. \tag{44}$$

But variational inequality (44) is precisely variational inequality (16) governing the supply chain network equilibrium with fixed demands.

Hence, we have the following result:

THEOREM 3. *A solution $(q^*, Q^{1*}, s^*, Q^{2*}) \in \mathcal{K}^4$ of the variational inequality (16) governing a supply chain network equilibrium coincides with the (via (30)–(33) and (34)–(37)) feasible link flow for the supernetwork $\mathcal{G}_S$ constructed above and satisfies variational inequality (24); equivalently, variational inequality (43). Hence, it is a transportation network equilibrium according to Theorem 2.*

We now describe how to recover the prices in the supply chain network with fixed demands. The vector of equilibrium prices ρ_3^* associated with the product at the demand markets can be obtained by setting (cf. (40) and (41)): $\rho_{3k}^* = C_{p_{ijj'k}} = \lambda_{w_k}^*$ for each demand market k. The vector of equilibrium prices ρ_2^* associated with retailers, in turn, can be obtained by setting (cf. (8) and (37)): $\rho_{2j}^* = \lambda_{w_k}^* - c_{a_{j'k}} = \left[\rho_{3k}^* - c_{jk}(Q^{2*})\right]$ for any j, k such that $q_{jk}^* > 0$. The equilibrium prices ρ_{1ij}^*, in turn, can be recovered by setting (cf. (3), (34), and (35)): $\rho_{1ij}^* = c_{a_i} + c_{a_{ij}} = \left[\frac{\partial f_i(Q^{1*})}{\partial q_{ij}} + \frac{\partial c_{ij}(q_{ij}^*)}{\partial q_{ij}}\right]$ for any i, j such that $q_{ij}^* > 0$.

We now further discuss the interpretation of the supply chain network equilibrium conditions. These conditions define the supply chain network equilibrium in terms of paths and path flows, which coincide with Wardrop's (1952) first principle of user-optimization in the context of transportation networks over the network given in Figure 2. Thus, we have an entirely new interpretation of supply chain network equilibrium in the case of known demands, which is as follows: all used paths connecting the source node 0 and a particular destination node have equal and minimal costs, and the cost on the utilized paths for this O/D pair is equal to the disutility (or the demand market price) that the consumers pay.

It is worth noting that the above identification yields and yet another application that can be formulated and solved as a transportation network equilibrium problem. For additional applications, including spatial price equilibrium problems and Walrasian price equilibrium problems, see Nagurney (1999) and the references therein.

We also point out that for a relatively price-insensitive product, such as, for example, gasoline or milk, the fixed demand assumption is, indeed, practical, and we expect that the model will provide a good approximation. We further emphasize that the equivalence established above between supply chain networks and transportation networks with fixed demands provides new opportunities for further modeling enhancements. In Section 5, we exploit this equivalence when we develop a dynamic supply chain network equilibrium model with time-varying demand.

5. Dynamic Supply Chain Networks with Time-Varying Demands

In this Section, we utilize the isomorphic transportation network established in Section 4 to develop a dynamic supply chain network model using an evolutionary variational inequality formulation. In Daniele, Maugeri, and Oettli (1998, 1999), evolutionary variational inequalities were utilized to model time-dependent transportation equilibria (see also Ran and Boyce (1996) and the references therein). Cojocaru, Daniele, and Nagurney (2005) demonstrated that time-dependent transportation equilibrium problems, as well as related dynamic spatial price equilibrium problems, and financial equilibrium problems, could be unified under a general evolutionary variational inequality definition over a unified constraint set. Nagurney, Liu, Cojocaru, and Daniele (2005) exploited the supernetwork equivalence between electric power networks and transportation networks, and developed an evolutionary variational inequality model for time-dependent electric power generation, distribution, and consumption. See the book by Nagurney (2006b) for additional theory and applications of dynamic supply chains.

In this Section, we consider the nonempty, convex, closed, bounded subset of the Hilbert space $L^2([0,T], R^Q)$ (where T denotes the time interval under consideration and $\mu = constant$ and very large) given by

$$\hat{\mathcal{K}} = \Big\{ x \in L^2([0,T], R^Q) : 0 \leq x(t) \leqslant \mu \text{ a.e. in } [0,T]; \sum_{p \in P_w} x_p(t) = d_w(t), \forall w, \text{ a.e. in } [0,T] \Big\}. \quad (45)$$

Hence, for definiteness, and greater ease in relating the discussion to the existing literature, we, without any loss of generality, consider the vector of path flows on the network at time t to be denoted by $x(t)$ with an individual element by $x_p(t)$ and with $d_w(t)$ denoting the demand associated with O/D pair w at time t.

Thus, we assume that the demands, $d_w(t)$, for all O/D pairs w are time-varying which means that the path flows will also change over time. We define:

$$\langle\langle \Phi, x \rangle\rangle = \int_0^T \langle \Phi(t), x(t) \rangle dt \quad (46)$$

where $\Phi \in L^2([0,T], R^Q)^*$ and $x \in L^2([0,T], R^Q)$. Let the function F be $F : \hat{\mathcal{K}} \to L^2([0,T], R^Q)$. We now provide the standardized form of the infinite-dimensional evolutionary (time-dependent) variational inequality (cf. Cojocaru, Daniele, and Nagurney (2005, 2006)): determine $x^* \in \hat{\mathcal{K}}$ such that:

$$\langle\langle F(x^*), x - x^* \rangle\rangle \geq 0, \quad \forall x \in \hat{\mathcal{K}}. \quad (47)$$

Sufficient conditions (including monotonicity-type conditions) that guarantee the existence of a solution to (47) are discussed in Daniele, Maugeri, and Oettli (1999).

Cojocaru, Daniele, and Nagurney (2006) have established that for the case of Hilbert spaces H (namely, $L^2([0,T], R^Q)$) the following infinite-dimensional projected dynamical systems (PDS) can be related to the evolutionary variational inequality (EVI) (47) as follows:

$$\frac{dx(t,\tau)}{d\tau} = \Pi_{\hat{\mathcal{K}}}(x(t,\tau), -F(x(t,\tau))), \quad x(t,0) \in \hat{\mathcal{K}}, \tag{48}$$

where

$$\Pi_{\hat{\mathcal{K}}}(y, -F(y)) = \lim_{\delta \to 0^+} \frac{P_{\hat{\mathcal{K}}}((y - \delta F(y)) - y)}{\delta}, \quad \forall y \in \hat{\mathcal{K}}, \tag{49}$$

with the projection operator $P_{\hat{\mathcal{K}}} : H \to \hat{\mathcal{K}}$ given by

$$\| P_{\hat{\mathcal{K}}}(z) - z \| = \inf_{y \in \hat{\mathcal{K}}} \| y - z \|. \tag{50}$$

Dupuis and Nagurney (1993) established the relationship between a projected dynamical system and a variational inequality in the case of finite dimensions. Cojocaru and Jonker (2004), subsequently, provided the relationship of the two formulations in infinite-dimensional Hilbert spaces. Recently, Cojocaru, Daniele, and Nagurney (2006) showed the following:

THEOREM 4. *Assume that $\hat{\mathcal{K}} \subseteq H$ is non-empty, closed, and convex. Assume also that $F : \hat{\mathcal{K}} \to H$ is a pseudo-monotone vector field, that is, for every pair of points $x, y \in \hat{\mathcal{K}}$, we have that*

$$\langle F(x), y - x \rangle \Longrightarrow \langle F(y), y - x \rangle \geq 0,$$

and that F is Lipschitz continuous, where H is a Hilbert space. Then the solutions of EVI (47) are the same as the critical points of the projected differential equation (48), that is, they are the functions $x^* \in \hat{\mathcal{K}}$ such that

$$\Pi_{\hat{\mathcal{K}}}(x^*(t), -F(x^*(t))) = 0, \tag{51}$$

and vice-versa.

Applying Theorem 4, we conclude that the solutions to the evolutionary variational inequality: determine $x^* \in \hat{\mathcal{K}}$ such that:

$$\int_0^T \langle F(x^*(t)), x(t) - x^*(t) \rangle dt \geq 0, \quad \forall x \in \hat{\mathcal{K}}, \tag{52}$$

coincide with the critical points of the equation:

$$\frac{dx(t,\tau)}{d\tau} = \Pi_{\hat{\mathcal{K}}}(x(t,\tau), -F(x(t,\tau))), \tag{53}$$

that is, the points satisfying

$$\Pi_{\hat{\mathcal{K}}}(x^*(t,\tau), -F(x^*(t,\tau))) \equiv 0 \text{ a.e. in } [0,T], \tag{54}$$

which are apparently stationary with respect to τ.

Note that in the formulation of the infinite-dimensional PDS (53), there are two "times," the meaning of which is discussed in Cojocaru, Daniele, and Nagurney (2007). Intuitively, at each moment $t \in [0,T]$, the solution of the evolutionary variational inequality (47) represents a static state of the underlying system. As t varies over $[0,T]$, the static states describe one (or more) curves of the equilibria. On the other hand, τ here is the time that describes the dynamics of the system until it reaches one of the equilibria of the curve.

The dynamic, evolutionary variational inequality analogue of the static, finite-dimensional variational inequality (23) is now immediate. We substitute the vector of path costs into (47) and we obtain the evolutionary variational inequality for time-dependent transportation network equilibria given by: determine $x^* \in \hat{\mathcal{K}}$ such that:

$$\langle\langle C(x^*), x - x^* \rangle\rangle \geq 0, \quad \forall x \in \hat{\mathcal{K}}, \tag{55}$$

where C is the vector of path costs.

According to Theorem 3, the supply chain network equilibrium problem with fixed demands can be reformulated as a fixed demand transportation network equilibrium problem over the supernetwork $\mathcal{G}_S$ given in Figure 2. Hence, the evolutionary variational inequality (55), in turn, provides us now with a dynamic version of the supply chain network model in which the demands vary over time, where the path costs are given by (38) and these are functions of path flows that now vary with time.

In the next Section, we illustrate the dynamic supply chain network model with concrete numerical examples.

6. Dynamic Numerical Supply Chain Network Examples with Computations

In this Section, we provide numerical examples in order to demonstrate how the theoretical results in this paper can be applied in practice. In particular, we consider numerical supply chain network examples with time-varying demands and product flows.

To solve the associated evolutionary variational inequality, we utilize the approach set forth in Cojocaru, Daniele, and Nagurney (2005, 2006, 2007), in which the time horizon T is discretized and at each fixed point in time we solve the associated projected dynamical system (cf. also Nagurney and Zhang (1996)). We have chosen the examples so that the corresponding vector field F satisfies the requirements in Theorem 4 (see also Nagurney, Dong, and Zhang (2002)), which we expect to be readily fulfilled in practice.

We utilized the Euler method for our numerical computations. The Euler method is induced by the general iterative scheme of Dupuis and Nagurney (1993) and has been applied by Nagurney and Zhang (1996) and Zhang and Nagurney (1997) to solve the variational inequality problem (23) in path flows as well as to approximate the continuous time trajectories associated with the corresponding projected dynamical system until the stationary point is attained. We applied the Euler method at discrete time points over the time interval T. Obviously, this procedure is correct if the continuity of the solution is guaranteed. Continuity results for solutions to evolutionary variational inequalities, in the case where $F(x(t)) = A(t)x(t) + B(t)$ is a linear operator, $A(t)$ is a continuous and positive definite matrix in $[0, T]$, and $B(t)$ is a continuous vector can be found in Barbagallo (2007). In the examples that we present here such assumptions are fulfilled. Of course, the examples could also be computed via the computational procedure given in Daniele, Maugeri, and Oettli (1999) but here we utilize a time-discretization approach which also has intuitive appeal.

The Euler method was implemented in FORTRAN and the computer system used was a Sun system at the University of Massachusetts at Amherst. The convergence criterion utilized was that the absolute value of the path flows between two successive iterations differed by no more than 10^{-5}. The sequence $\{\alpha_\tau\}$ in the Euler method (cf. Nagurney and Zhang (1996)) was set to: $.1\{1, \frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \ldots\}$. The Euler method was initialized by distributing the demand for each O/D pair equally among the paths connecting the respective O/D pair for each discretized point in time.

Example 1

In the first numerical example, the supply chain network consisted of one manufacturer, three retailers, and one demand market as depicted in Figure 3. The supernetwork representation which allows for the transformation (as proved in Section 4) to a transportation network equilibrium problem is given also in Figure 3. Hence, in the first numerical example (see also Figure 2) we had that: $m = 1$, $n = 3$, $n' = 3'$, and $o = 1$.

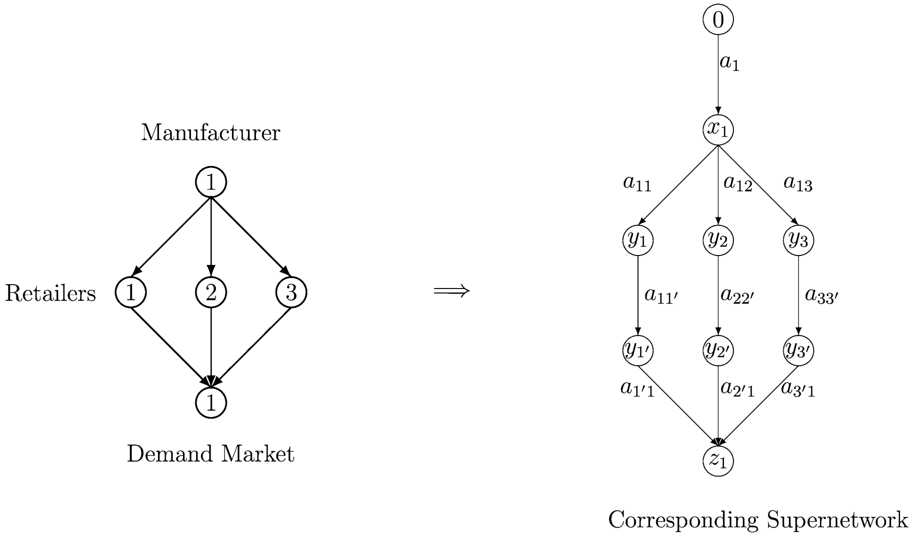


Figure 3. Supply Chain Network and Corresponding Supernetwork $\mathcal{G}_{\mathcal{S}}$ for Numerical Examples 1 and 2.

The notation is presented here in the form of the supply chain network model as delineated in Table 1. We provide the complete supernetwork representation in terms of O/D pairs, paths, etc. The translations of the equilibrium path flows, link flows, and travel disutilities into the equilibrium flows and prices is then given, for completeness, and easy reference.

The production cost function for the manufacturer was given by:

$$f_1(q(t)) = 2.5q_1(t)^2 + 2q_1(t).$$

The transaction cost functions faced by the manufacturer and associated with transacting with the retailers were given by:

$$c_{11}(q_{11}(t)) = .5q_{11}(t)^2 + 3.5q_{11}(t), \quad c_{12}(q_{12}(t)) = .5q_{12}(t)^2 + 2.5q_{12}(t),$$
$$c_{13}(q_{13}(t)) = .5q_{13}(t)^2 + 1.5q_{13}(t).$$

The operating costs of the retailers, in turn, were given by:

$$c_1(Q^1(t)) = .5(q_{11}(t))^2, \ c_2(Q^1(t)) = .5(q_{12}(t))^2, \ c_3(Q^1(t)) = .5(q_{13}(t))^2.$$

The unit transaction costs associated with transacting between the retailers and the demand market were:

$$c_{11}(Q^2(t)) = q_{11}(t) + 1, \ c_{21}(Q^2(t)) = q_{21}(t) + 5, \ c_{31}(Q^2(t)) = q_{31}(t) + 10.$$

We utilized the supernetwork representation of this example depicted in Figure 3 with the links enumerated as in Figure 3 in order to solve the problem via the Euler method. Note that there are 9 nodes and 10 links in the supernetwork in Figure 3. Using the procedure outlined in Section 4, we defined O/D pair $w_1 = (0, z_1)$ with the user link travel cost functions as given in (34)–(37).

There were three paths in P_{w_1} denoted by: p_1, p_2, p_3. The paths were comprised of the following links:

$$p_1 = (a_1, a_{11}, a_{11'}, a_{1'1}), \quad p_2 = (a_1, a_{12}, a_{22'}, a_{2'1}), \quad p_3 = (a_1, a_{13}, a_{33'}, a_{3'1}).$$

The time horizon $T = 1$. The time-varying demand function was given by:

$$d_1(t) = 100 + 10t.$$

We discretized the time horizon T as follows: $t_0 = 0$, $t_1 = \frac{1}{2}$, and $t_2 = T = 1$. We report the solutions obtained by the Euler method at each discrete time step, for which we had, respectively, demands: $d_1(t_0) = 100$; $d_1(t_1) = 105$, and $d_1(T) = 110$.

Example 1: Solution at Time $t = t_0 = 0$:

The Euler method converged and yielded the following equilibrium path flow pattern:

$$x^*_{p_1}(t_0) = 34.44, \quad x^*_{p_2}(t_0) = 33.44, \quad x^*_{p_3}(t_0) = 32.12.$$

The corresponding equilibrium link flows (cf. also the supernetwork in Figure 3) were:

$$f^*_{a_1}(t_0) = 100.00,$$

$$f^*_{a_{11}}(t_0) = 34.44, \quad f^*_{a_{12}}(t_0) = 33.44, \quad f^*_{a_{13}}(t_0) = 32.12,$$

$$f^*_{a_{11'}}(t_0) = 34.44, \quad f^*_{a_{22'}}(t_0) = 33.44, \quad f^*_{a_{33'}}(t_0) = 32.12,$$

$$f^*_{a_{1'1}}(t_0) = 34.44, \quad f^*_{a_{2'1}}(t_0) = 33.44, \quad f^*_{a_{3'1}}(t_0) = 32.12.$$

The incurred equilibrium path travel costs (cf. (38)) were: $C_{p_1}(t_0) = C_{p_2}(t_0) = C_{p_3}(t_0) = \lambda^*_{w_1}(t_0) = 609.83$.

We now provide the translations of the above equilibrium flows into the supply chain network flow and price notation using (30), (31), (32), and (33).

The flows were:

$$Q^{1*}(t_0) := q^*_{11}(t_0) = 34.44, \quad q^*_{12}(t_0) = 33.44, \quad q^*_{13}(t_0) = 32.12,$$
$$s^*_1(t_0) = 34.44, \quad s^*_2(t_0) = 33.44, \quad s^*_3(t_0) = 32.12,$$
$$Q^{2*}(t_0) := q^*_{11}(t_0) = 34.44, \quad q^*_{21}(t_0) = 33.44, \quad q^*_{31}(t_0) = 32.12,$$

and the production quantity was: $q^*_1(t_0) = 100$.

The demand price at the demand market was, hence, (cf. (40)):

$$\rho^*_{31}(t_0) = 609.83,$$

which corresponds to the travel costs on the paths (all are used) connecting the O/D pair.

It is easy to verify that the equilibrium conditions were satisfied with excellent accuracy.

Example 1: Solution at Time $t = t_1 = \frac{1}{2}$:

The Euler method converged and yielded the following equilibrium path flow pattern:

$$x^*_{p_1}(t_1) = 36.11, \quad x^*_{p_2}(t_1) = 35.11, \quad x^*_{p_3}(t_1) = 33.78.$$

The corresponding equilibrium link flows (cf. also the supernetwork in Figure 3) were:

$$f^*_{a_1}(t_1) = 105.00,$$
$$f^*_{a_{11}}(t_1) = 36.11, \quad f^*_{a_{12}}(t_1) = 35.11, \quad f^*_{a_{13}}(t_1) = 33.78,$$
$$f^*_{a_{11'}}(t_1) = 36.11, \quad f^*_{a_{22'}}(t_1) = 35.11, \quad f^*_{a_{33'}}(t_1) = 33.78,$$
$$f^*_{a_{1'1}}(t_1) = 36.11, \quad f^*_{a_{2'1}}(t_1) = 35.11, \quad f^*_{a_{3'1}}(t_1) = 33.78,$$

with a production quantity: $q^*_1(t_1) = 105$. The equilibrium path travel costs were now: $C_{p_1}(t_1) = C_{p_2}(t_1) = C_{p_3}(t_1) = \lambda^*_{w_1}(t_1) = 639.83$.

The translations into the corresponding equilibrium supply chain flows at time t_1 can easily be done as described for time t_0.

The demand price at the demand market was now:

$$\rho^*_{31}(t_1) = 639.83,$$

which corresponds to the travel costs on the paths (all paths are again used) connecting the O/D pair.

It is easy to verify that the equilibrium conditions were again satisfied with excellent accuracy.

Example 1: Solution at Time $t = T = 1$:

We applied the Euler method to the end of the time horizon where $T = 1$. The Euler method now yielded the following equilibrium path flow pattern:

$$x^*_{p_1}(T) = 37.77, \quad x^*_{p_2}(T) = 36.77, \quad x^*_{p_3}(T) = 35.45.$$

The corresponding equilibrium link flows (cf. also the supernetwork in Figure 3) were:

$$f^*_{a_1}(T) = 110.00,$$
$$f^*_{a_{11}}(T) = 37.77, \quad f^*_{a_{12}}(T) = 36.77, \quad f^*_{a_{13}}(T) = 35.45,$$
$$f^*_{a_{11'}}(T) = 37.77, \quad f^*_{a_{22'}}(T) = 36.77, \quad f^*_{a_{33'}}(T) = 35.45,$$
$$f^*_{a_{1'1}}(T) = 37.77, \quad f^*_{a_{2'1}}(T) = 36.77, \quad f^*_{a_{3'1}}(T) = 35.45.$$

The translations into the corresponding equilibrium supply chain flows can be easily done as above for time t_0.

The demand price at the demand market was now:

$$\rho^*_{31}(T) = 669.83,$$

which is equal to $\lambda^*_{w_1}(T) = C_{p_1}(T) = C_{p_2}(T) = C_{p_3}(T)$.

Explicit Formulae for Example 1 for the Time-Dependent Equilibria

We now note that, due to the linearity of F in this example, as well as the separability of the components of F, and the special structure of the topology of the supernetwork in Figure 3, we can write down explicit formulae for the path flows over time $[0, T]$. See also, Dafermos and Sparrow (1969) who made the same observation in the context of transportation network equilibrium problems on networks in which all paths connecting an O/D pair consisted of single links, and the user link cost functions were linear and separable. Cojocaru, Daniele, and Nagurney (2005, 2006) provided explicit formulae for solutions to dynamic transportation network examples of such special topologies and cost structures.

In particular, we obtain the following formulae for the equilibrium path flows for Example 1 at each point t:

$$x^*_{p_1}(t) = 3.33t + 34.44,$$

$$x^*_{p_2}(t) = 3.33t + 33.44,$$
$$x^*_{p_3}(t) = 3.33t + 32.12,$$

and these formulae are valid even for $T > 1$, that is, outside the range [0, 1], which is of concern here. We also have an explicit formula for the travel disutility where:

$$\lambda^*_{w_1}(t) = 60t + 609.83, \quad \text{for} \quad t \in [0, T].$$

We now, for completeness, translate these formulae into supply chain network model formulae (30)–(33) with time-varying flows (see also (55)). Please refer also to the supernetwork in Figure 3. In particular, we have the time-dependent equilibrium supply chain flows are given by:

$$\begin{aligned}
q^*_1(t) &= f^*_{a_1}(t) = x^*_{p_1}(t) + x^*_{p_2}(t) + x^*_{p_3}(t) = 10t + 100;\\
Q^{1*}(t) := q^*_{11}(t) &= f^*_{a_{11}}(t) = x^*_{p_1}(t) = 3.33t + 34.33,\\
q^*_{12}(t) &= f^*_{a_{12}}(t) = x^*_{p_2}(t) = 3.33t + 33.44,\\
q^*_{13}(t) &= f^*_{a_{13}}(t) = x^*_{p_3}(t) = 3.33t + 32.12;\\
s^*_1(t) &= f^*_{a_{11'}}(t) = x^*_{p_1}(t) = 3.33t + 34.44,\\
s^*_2(t) &= f^*_{22'}(t) = x^*_{p_2}(t) = 3.33t + 33.44,\\
s^*_3(t) &= f^*_{a_{33'}}(t) = x^*_{p_3}(t) = 3.33t + 32.12,
\end{aligned}$$

and

$$\begin{aligned}
Q^{2*} := q^*_{11}(t) &= f^*_{a_{1'1}}(t) = x^*_{p_1}(t) = 3.33t + 34.44,\\
q^*_{21}(t) &= f^*_{a_{2'1}}(t) = x^*_{p_2}(t) = 3.33t + 33.44,\\
q^*_{31}(t) &= f^*_{a_{3'1}}(t) = x^*_{p_3}(t) = 3.33t + 32.12.
\end{aligned}$$

Example 2: A Numerical Supply Chain Example with Step-wise Time Varying Demand

The second example had the same data as Example 1, except that the demand now had a step-wise structure. The supply chain network and the supernetwork were, hence, as in Figure 3. In particular, the demand was of the form given below on the time interval $[0, T]$:

$$d_1(t) = \begin{cases} s_1, & \text{if} \quad 0 < t \le t_1, \\ s_2, & \text{if} \quad t_1 < t \le t_2, \\ \dots, & \\ s_k, & \text{if} \quad t_{k-1} < t \le t_k = T, \end{cases}$$

where, in this example, we have that:

$$d_1(t) = \begin{cases} 100, \text{ if} & 0 < t \le t_1 = \frac{1}{2}, \\ 110, \text{ if} & t_1 < t \le t_2 = T = 1. \end{cases}$$

Such a structure may reflect, for example, a seasonable demand for a product

In this setting, we know that the equilibrium curve (solution of the evolutionary variational inequality) is a step function, with the steps given by the function $d_1(t)$, where:

$$x^*(t) = \begin{cases} x_1^*, \text{ if} & 0 < t \le t_1 = \frac{1}{2}, \\ x_2^*, \text{ if} & t_1 < t \le t_2 = 1 = T. \end{cases}$$

Again, given the simplicity of the supernetwork topology and the cost structure, we obtain an explicit solution:

$$\begin{aligned} x^*(t) &= (x^*_{p_1}(t), x^*_{p_2}(t), x^*_{p_3}(t)) \\ &= \begin{cases} (34.44, 33.44, 32.12), \text{ if} & 0 < t \le t_1 = \frac{1}{2}, \\ (37.77, 36.77, 35.45), \text{ if} & t_1 < t \le t_2 = 1 = T. \end{cases} \end{aligned}$$

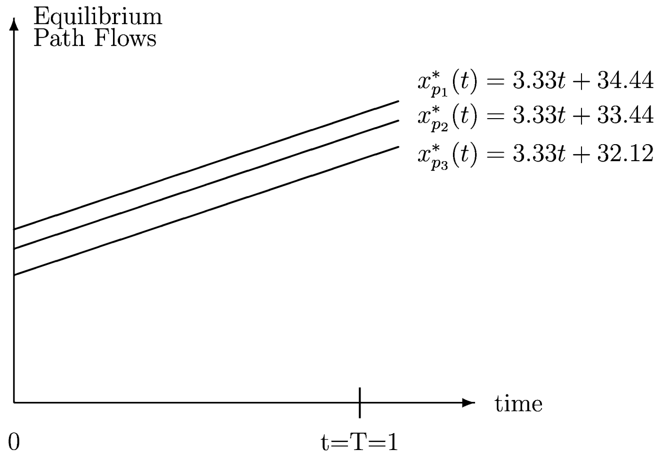


Figure 4. Time-Dependent Equilibrium Path Flows for Example 1.

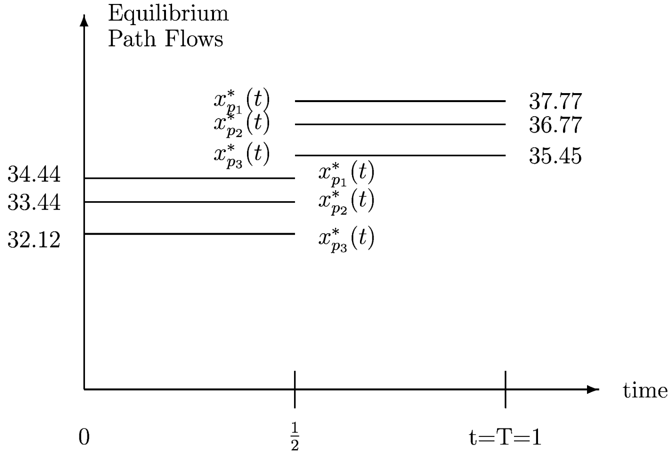


Figure 5. Time-Dependent Equilibrium Path Flows for Example 2 with Step-Wise Demand.

Of course, the transformation of the these equilibrium path flows into the equilibrium link flows and the supply chain network flows can be done as was done for Example 1 since the supernetwork topology is one and the same for Examples 1 and 2.

In Figures 4 and 5 we provide the graphs of the time-dependent equilibrium path flows for Examples 1 and 2, respectively.

Example 3

In the third numerical example, the supply chain network consisted of two manufacturers, one retailer, and two demand markets. Hence, we now had that $m = 2$, $n = 1$, $n' = 1'$, and $o = 2$.

The data were now as follows: The production cost functions for the manufacturers were given by:

$$f_1(q(t)) = 2.5q_1(t)^2 + q_1(t)q_2(t) + 2q_1(t),$$
$$f_2(q(t)) = 2.5q_2(t)^2 + q_2(t)q_1(t) + 2q_2(t).$$

The transaction cost functions faced by the manufacturers and associated with transacting with the retailers were given by:

$$c_{11}(q_{11}(t)) = .5q_{11}(t)^2 + 3.5q_{11}(t), \quad c_{21}(q_{21}(t)) = .5q_{21}(t)^2 + 1.5q_{21}(t).$$

The operating cost of the retailer, in turn, was given by:

$$c_1(Q^1(t)) = .5(q_{11}(t))^2.$$

The unit transaction costs associated with transacting between the retailers and the demand market were:

$$c_{jk}(Q^2(t)) = q_{jk}(t) + 1, \quad \text{for } j = 1,2; k = 1,2.$$

We utilized the supernetwork representation of this example depicted in Figure 6 with the links enumerated as in Figure 6 in order to solve the problem via the Euler method. Note that there are 7 nodes and 7 links in the supernetwork in Figure 6. Using the procedure outlined in Section 4, we defined O/D pair $w_1 = (0, z_1)$ and O/D pair $w_2 = (0, z_2)$ with the user link travel cost functions as given in (40)–(43).

There were two paths in P_{w_1} denoted by: p_1, p_2 and two paths in P_{w_2} denoted by: p_3 and p_4, respectively. The paths were comprised of the following links:

$$p_1 = (a_1, a_{11}, a_{11'}, a_{1'1}), \quad p_2 = (a_2, a_{21}, a_{11'}, a_{1'1}),$$
$$p_3 = (a_1, a_{11}, a_{11'}, a_{1'2}), \quad p_4 = (a_2, a_{21}, a_{11'}, a_{1'2}).$$

The time horizon $T = 1$. The time-varying demand functions were given by:

$$d_1(t) = 100 + 5t, \quad d_2(t) = 80 + 4t.$$

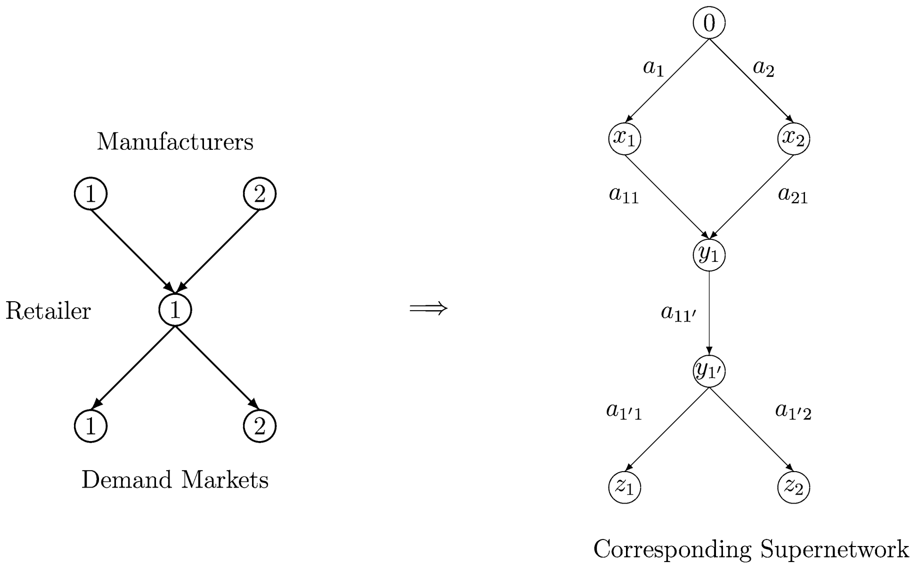


Figure 6. Supply Chain Network and Corresponding Supernetwork $\mathcal{G}_{\mathcal{S}}$ for Numerical Example 3.

We discretized the time horizon T as follows: $t_0 = 0$, $t_1 = \frac{1}{2}$, and $t_2 = T = 1$. We report the solutions obtained by the Euler method at each discrete time step, for which we had, respectively, demands: $d_1(t_0) = 100$, $d_1(t_1) = 102.5$, and $d_1(T) = 105$, and $d_2(t_0) = 80$, $d_2(t_1) = 82$, and $d_2(T) = 84$.

Example 3: Solution at Time $t = t_0 = 0$:

We applied the Euler method to the beginning of the time horizon where $t = t_0 = 0$. The Euler method now yielded the following equilibrium path flow pattern:

$$x^*_{p_1}(t_0) = 49.90, \quad x^*_{p_2}(t_0) = 50.10, \quad x^*_{p_3}(t_0) = 39.90, \quad x^*_{p_4}(t_0) = 40.10.$$

The corresponding equilibrium link flows (cf. also the supernetwork in Figure 4) were:

$$f^*_{a_1}(t_0) = 89.80, \quad f^*_{a_2}(t_0) = 90.20$$
$$f^*_{a_{11}}(t_0) = 89.80, \quad f^*_{a_{21}}(t_0) = 90.20,$$
$$f^*_{a_{11'}}(t_0) = 180.00,$$
$$f^*_{a_{1'1}}(t_0) = 100.00, \quad f^*_{a_{1'2}}(t_0) = 80.00,$$

with incurred equilibrium path travel costs: $C_{p_1}(t_0) = C_{p_2}(t_0) = \lambda^*_{w_1}(t_0) = 815.50$ and $C_{p_3}(t_0) = C_{p_4}(t_0) = \lambda^*_{w_2}(t_0) = 815.50$.

The translations into the corresponding equilibrium flows are now given:

$$Q^{1*}(t_0) := q^*_{11}(t_0) = 89.80, \quad q^*_{21}(t_0) = 92.90,$$
$$s^*_1(t_0) = f^*_{a_{11'}}(t_0) = 180.00,$$
$$Q^{2*}(t_0) := q^*_{11}(t_0) = 100.00, \quad q^*_{12}(t_0) = 80.00.$$

The demand prices at the demand markets were:

$$\rho^*_{31}(t_0) = 815.50, \quad \rho^*_{32}(t_0) = 815.50,$$

which correspond to the travel costs on the paths (all are used) connecting the respective O/D pair.

Example 3: Solution at Time $t = t_1 = \frac{1}{2}$:

We applied the Euler method to time $t = t_1 = \frac{1}{2}$. The Euler method now yielded the following equilibrium path flow pattern:

$$x^*_{p_1}(t_1) = 51.15, \quad x^*_{p_2}(t_1) = 51.35, \quad x^*_{p_3}(t_1) = 40.90, \quad x^*_{p_4}(t_1) = 41.10.$$

The corresponding equilibrium link flows (cf. also the supernetwork in Figure 3) were:

$$f^*_{a_1}(t_1) = 92.05, \quad f^*_{a_2}(t_1) = 92.45$$
$$f^*_{a_{11}}(t_1) = 92.05, \quad f^*_{a_{21}}(t_1) = 92.45,$$
$$f^*_{a_{11'}}(t_1) = 184.50,$$
$$f^*_{a_{1'1}}(t_1) = 102.50, \quad f^*_{a_{1'2}}(t_1) = 82.00,$$

with equilibrium path costs:

$$C_{p_1}(t_1) = C_{p_2}(t_1) = \lambda^*_{w_1}(t_1) = 835.75,$$
$$C_{p_3}(t_1) = C_{p_4}(t_1) = \lambda^*_{w_2}(t_1) = 835.75.$$

The translations into the corresponding equilibrium supply chain flows are now given:

$$Q^{1*}(t_1) := q^*_{11}(t_1) = 92.05, \quad q^*_{21}(t_1) = 92.45,$$
$$s^*_1(t_1) = f^*_{a_{11'}}(t_1) = 184.50,$$
$$Q^{2*}(t_1) := q^*_{11}(t_1) = 102.50, \quad q^*_{12}(t_1) = 82.00.$$

The demand prices at the demand markets were now:

$$\rho^*_{31}(t_1) = 835.75, \quad \rho^*_{32}(t_1) = 835.75,$$

which correspond to the travel costs on the paths (all are used) connecting the respective O/D pair.

Example 3: Solution at Time $t = T = 1$:

Finally, we applied the Euler method to the end of the time horizon where $t = T$. The Euler method now yielded the following equilibrium path flow pattern:

$$x^*_{p_1}(T) = 52.40, \quad x^*_{p_2}(T) = 52.60, \quad x^*_{p_3}(T) = 41.90, \quad x^*_{p_4}(T) = 42.10.$$

The corresponding equilibrium link flows (cf. Figure 4) were:

$$f^*_{a_1}(T) = 94.30, \quad f^*_{a_2}(T) = 94.70,$$
$$f^*_{a_{11}}(T) = 94.30, \quad f^*_{a_{21}}(T) = 94.70,$$

$$f^*_{a_{11'}}(T) = 189.00,$$

$$f^*_{a_{1'1}}(T) = 105.00, \quad f^*_{a_{1'2}}(T) = 84.00.$$

The equilibrium path costs, in turn, were now:

$$C_{p_1}(T) = C_{p_2}(T) = \lambda^*_{w_1}(T) = 856.00, \;\; C_{p_3}(T) = C_{p_4}(T) = \lambda^*_{w_2}(T) = 856.00.$$

The translations into the corresponding equilibrium supply chain flows were, hence:

$$Q^{1*}(T) := q^*_{11}(T) = 94.30, \quad q^*_{21}(T) = 94.70,$$

$$s^*_1(T) = f^*_{a_{11'}}(T) = 189.00,$$

$$Q^{2*}(T) := q^*_{11}(T) = 105.00, \quad q^*_{12}(T) = 84.00.$$

The demand prices at the demand markets were now:

$$\rho^*_{31}(T) = 856.00, \quad \rho^*_{32}(T) = 856.00,$$

which correspond to the travel costs on the paths (all are used) connecting the respective O/D pair.

Acknowledgments

This research was supported by NSF Grant No. IIS-002647. The first author also acknowledges support from the Radcliffe Institute for Advanced Study at Harvard University under its 2005–2006 Felowship Program. The support provided is gratefully acknowledged.

The authors would also like to thank Terry L. Friesz for organizing and hosting the NSF-sponsored workshop at Pennsylvania State University, May 9–11, 2005, on Network Science, Nonlinear Science and Infrastructure Systems at which an earlier version of this paper was presented. Finally, thanks are extended to the participants for stimulating discussions at the workshop.

References

Barbagallo, A., 2007. Regularity results for time-dependent variational and quasi-variational inequalities and application to the calculation of dynamic traffic network. *Applied Sciences* 17, 277–304.

Bazaraa, M. S., Sherali, H. D., Shetty, C. M., 1993. Nonlinear Programming: Theory and Algorithms. John Wiley & Sons, New York.

Beckmann, M. J., McGuire, C. B., Winsten, C. B., 1956. Studies in the Economics of Transportation. Yale University Press, New Haven, Connecticut.

Boyce, D. E., Mahmassani, H. S., Nagurney, A., 2005. A retrospective on Beckmann, McGuire, and Winsten's Studies in the Economics of Transportation. Papers in Regional Science 84, 85–103.

Cojocaru, M.-G., Daniele, P., Nagurney, A., 2005. Projected dynamical systems and evolutionary variational inequalities via Hilbert spaces with applications. Journal of Optimization Theory and Applications 27(3), 1–15.

Cojocaru, M.-G., Daniele, P., Nagurney, A., 2006. Double-layered dynamics: A unified theory of projected dynamical systems and evolutionary variational inequalities. *European Journal of Operational Research*, 175, 494–507.

Cojocaru, M.-G., Daniele, P., Nagurney, A. 2007. Projected dynamical systems, evolutionary variational inequalities, applications, and a computational procedure. Pareto Optimality, Game Theory and Equilibria. A. Migdalas, P. M. Pardalos, and L. Pitsoulis, editors, Springer Verlag, in press.

Cojocaru M.-G., Jonker, L. B., 2004. Existence of solutions to projected differential equations in Hilbert spaces. Proceedings of the American Mathematical Society 132, 183–193.

Cournot, A. A., 1838. Researches into the Mathematical Principles of the Theory of Wealth. English translation, MacMillan, England.

Dafermos, S., 1980. Traffic equilibrium and variational inequalities. Transportation Science 14, 42–54.

Dafermos, S., 1982. The general multimodal network equilibrium problem with elastic demand. Networks 12, 57–72.

Dafermos, S., Nagurney, A., 1984. Stability and sensitivity analysis for the general network equilibrium – travel choice model. In: Volmuller, J., Hamerslag, R. (Eds.), Proceedings of the Ninth International Symposium on Transportation and Traffic Theory. VNU Science Press, Utrecht, The Netherlands, pp. 217–232.

Dafermos, S., Nagurney, A., 1987. Oligopolistic and competitive behavior of spatially separated markets. Regional Science and Urban Economics 17, 245–254.

Dafermos, S. C., Sparrow, F. T., 1969. The traffic assignment problem for a general network. Journal of Research of the National Bureau of Standards 73B, 91–118.

Daniele, P., Maugeri, A., Oettli, W., 1998. Variational inequalities and time-dependent traffic equilibria. Comptes Rendue Academie des Science, Paris 326, serie I, 1059–1062.

Daniele, P., Maugeri, A., Oettli, W., 1999. Time-dependent traffic equilibria. Journal of Optimization Theory and its Applications 103, 543–555.

Dupuis, P., Nagurney, A., 1993. Dynamical systems and variational inequalities. Annals of Operations Research 44, 9–42.

Fisk, C., Boyce, D. E., 1983. Alternative variational inequality formulation of the network equilibrium-travel choice problem. Transportation Science 17, 454–463.

Gabay, D., Moulin, H., 1980. On the uniqueness and stability of Nash equilibria in noncooperative games. In: Applied Stochastic Control in Econometrics and Management Science, Bensoussan, A., Kleindorfer, P., Tapiero, C. S. (Eds.), North-Holland, Amsterdam, The Netherlands, 271–294.

Kinderlehrer, D., Stampacchia, G., 1980. An Introduction to Variational Inequalities and Their Applications. Academic Press, New York.

Nagurney, A., 1999. Network Economics: A Variational Inequality Approach, second and revised edition. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Nagurney, A. 2006. On the relationship between supply chain and transportation network equilibria: A supernetwork equivalence with computations. Transportation Research E 42, 293–316.

Nagurney, A. 2006b. Supply Chain Network Economics: Dynamics of Prices, Flows, and Profits, Edward Elgar Publishing, Cheltenham, England.

Nagurney, A., Dong, J., 2002. Supernetworks: Decision-Making for the Information Age. Edward Elgar Publishing, Cheltenham, England.

Nagurney, A., Dong, J., Zhang, D., 2002. A supply chain network equilibrium model. Transportation Research E 38, 281–303.

Nagurney, A., Liu, Z., Cojocaru, M., Daniele, P., 2005. Static and dynamic transportation network equilibrium reformulations of electric power supply chain networks with known demands, to appear in Transportation Research E.

Nagurney, A., Zhang, D., 1996. Projected Dynamical Systems and Variational Inequalities with Applications. Kluwer Academic Publishers, Boston, Massachusetts.

Nagurney, A., Zhang, D., 1997. Projected dynamical systems in the formulation, stability analysis, and computation of fixed demand traffic network equilibria. Transportation Science 31, 147–158.

Nash, J. F., 1950. Equilibrium points in n-person games. Proceedings of the National Academy of Sciences 36, 48–49.

Nash, J. F., 1951. Noncooperative games. Annals of Mathematics 54, 286–298.

Ran, B., Boyce, D. E., 1996. Modeling Dynamic Transportation Networks. Springer-Verlag, Berlin, Garmany.

Smith, M. J., 1979. Existence, uniqueness, and stability of traffic equilibria. Transportation Research 13B, 259–304.

Wardrop, J. G., 1952. Some theoretical aspects of road traffic research. In: Proceedings of the Institution of Civil Engineers, Part II 1, 325–378.

Zhang, D., Nagurney, A., 1997. Formulation, stability, and computation of traffic network equilibria as projected dynamical systems. Journal of Optimization Theory and Applications 93, 417–444.

Zhang, P., Peeta, S., Friesz, T. L., 2005. Dynamic game theoretic model of multi-layer infrastructure networks. *Networks and Spatial Economics* 5, 146–178.

Chapter 13

Some Amazing Properties of Road Traffic Network Equilibria

Hillel Bar-Gera[1] and David Boyce[2]

[1] *Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel, E-mail: bargera@bgu.ac.il*

[2] *Department of Civil and Environmental Engineering, Northwestern University, Evanston, IL 60208, E-mail: d-boyce@northwestern.edu*

Abstract One of the first mathematical models of a physical network interacting with human behavior was the model of road traffic equilibria with variable flow (demand) formulated by Martin Beckmann and colleagues in 1954. Beckmann applied the recently-proved theorem of Kuhn and Tucker to incorporate an assumption and two hypotheses concerning road traffic into a single mathematical formulation. The model considers a road network consisting of nodes and links. Associated with each directional link is an increasing function relating its travel time, or generalized travel cost, to its flow. The behavioral hypotheses represented by the model are as follows:

1. All used routes from node p to node q have equal travel times, and no unused route has a lower travel time;
2. The total flow over all routes from node p to node q is determined by a decreasing function of this minimum and equal, or equilibrium, travel time.

In large-scale implementations of the model, nodes p and q represent small areas called zones, at which flows originate and terminate; other nodes represent intersections on the road network. The formulation minimizes an artificial function, subject to definitional constraints. The optimality conditions of this model correspond to the above two hypotheses. Subsequently, more general formulations were investigated based on variational inequality, nonlinear complementarity and fixed point theory.

Beckmann's formulation and its descendents considered traffic flows over a relatively long period of time, during which network conditions may be regarded as constant. The peak commuting period in the morning or evening is a typical example. Such models are static, and the flows departing from and arriving at nodes are constant over the time period. Models that consider shorter periods of time, and for which the departure and arrival rates are variables, are dynamic. These models seek to represent the effect of changing network conditions during a longer time period, including accidents and other incidents disrupting flow.

Although Beckmann did not propose an algorithm for solving his formulation, in the 1970s researchers began to solve large-scale traffic equilibria. Until recently, these solutions were rather approximate, and did not reveal the structure of the solution, especially with regard to the number and pattern of equilibrium routes. In 2003, Bar-Gera and Boyce proposed an algorithm that reveals this structure for the first time. Subsequently, they began to explore the properties of this solution for large-scale implementations, such as for the Chicago region. The initial results of these explorations for the Chicago region were unexpected and regarded as "astonishing" by one informed observer. One result examined is the relation between the number of routes between a pair of zones and the frequency with which this number occurs in the network. The authors observed that the number of routes increases greatly as the level of congestion increases.

This chapter seeks to introduce traffic network equilibrium models to scholars from a broad range of backgrounds, mainly focusing on static models of urban road traffic. Findings on the solution properties of static models for a large network for three congestion levels are presented. A discussion of the applicability of the findings to other types of networks, such as electrical power and supply chain networks, concludes the paper.

Keywords: traffic equilibria; congestion

1. Introduction

Perhaps the first mathematical model of a functioning, physical network and associated human behavior represented road traffic equilibria with variable flow (demand), as formulated by Martin Beckmann in 1954 (Beckmann et al., 1956). Beckmann applied the recently-proved theorem of Kuhn and Tucker (1951) to incorporate an assumption and two hypotheses concerning road traffic into a single mathematical formulation. The model considers a road network consisting of nodes and links. Associated with each directional link is a function relating its travel time, or generalized travel cost, to its flow. As represented by the function, travel time is assumed to increase as flow increases without limit. The behavioral hypotheses represented by the model are the following:

1. All used routes from node p to node q have equal and minimal travel times;
2. The total flow over all routes from node p to node q is a decreasing function of this minimum and equal, or equilibrium, travel time, as well as exogenous originating and terminating total flows.

In large-scale implementations of the model, nodes p and q represent small areas called zones, at which the flows originate and terminate; other nodes represent intersections on the road network.

The formulation that Beckmann devised may be described as the minimization of an artificial function, subject to definitional constraints. The

Kuhn-Tucker optimality conditions of this model correspond to the above two hypotheses. Subsequently, more general formulations were investigated based on variational inequality, nonlinear complementarity and fixed point theory; see for example Nagurney (1999) and Bar-Gera and Boyce (2003).

Beckmann's formulation and its descendents considered traffic flows over a relatively long period of time, during which network conditions may be regarded as constant. The peak commuting period in the morning or evening is a typical example. Such models are termed *static*, and the flows departing from and arriving at nodes are constant over the time period. Models that consider shorter periods of time, and for which the departure and arrival rates are variables, are termed *dynamic*. These models seek to represent the effect of changing network conditions during a longer time period, including accidents and other incidents disrupting flow.

Beckmann did not propose an algorithm for solving his formulation. Given the primitive computers of the time, he and his coauthors did not consider its solution for large-scale urban networks as being possible (Boyce, 2007). Subsequently, many researchers have taken up the problem of solving models of large-scale traffic equilibria. Until recently, these solutions were rather approximate, and did not reveal the structure of the solution, especially with regard to the number and pattern of equilibrium routes. Bar-Gera (2002, 2006) and Bar-Gera and Boyce (2003) proposed an algorithm that reveals this structure for the first time. Subsequently, they began to explore the properties of this solution for large-scale implementations, such as the 1790 zone, 39,000 link model of the Chicago region (Bar-Gera and Boyce, 2005).

The objective of this paper is to present new findings about the structure of the solution of such models, and to interpret these findings. The paper consists of the following sections:

1. a relatively non-technical description of the traffic network equilibrium model;
2. presentation of findings on the solution properties of traffic equilibrium models for a large network for three congestion levels;
3. a brief review of models of other types of networks, such as freight transportation, supply chain, utility and data telecommunication networks, and comments on the applicability of these findings for road transportation to such networks.

2. Traffic Network Equilibrium Model

The traffic network equilibrium problem stated in the Introduction may be formally expressed in several mathematical forms. One of the most general is the fixed point formulation. The variational inequality problem formulation is also quite general, which is closely related to the nonlinear complementarity problem. Formulations based on nonlinear optimization with linear constraints are more restricted, but more amendable to solution. Finally, newer developments based on orthogonality and normal cones by Patriksson and Rockafellar (2003) offer perhaps the most general synthesis to this point; see also Patriksson (2006).

The following description depicts the model in terms of its basic variables and assumptions; a more detailed description including the solution procedure is found in the Appendix. Consider a study area that is divided into a set of zones, Z, connected by transit services and by a road network. The road network consists of a set of nodes, N, and a set of directional links, A. A route is a sequence of nodes, $[v_1, \ldots, v_k]$, such that $[v_i, v_{i+1}] \in A$. The set of available routes from origin $p \in Z$ to destination $q \in Z$ is R_{pq}, and the set of all routes is R.

The purpose of the model is to predict:

1. the mode-origin-destination (MOD) flow, d_{mpq}, in persons per hour, for every mode $m \in$ {a=auto, t=transit}, origin $p \in Z$, and destination $q \in Z$;
2. the distribution of auto OD flows to route flows h_r, for every route $r \in R$, as determined by the route choices of travelers.

Auto OD flows are the sum of person-trips per hour divided by a constant auto occupancy factor (1.2 persons/auto) and the given OD truck flows, d_{pq}^{truck}, in equivalent passenger cars per hour: $d_{pq}^{auto} = d_{apq}/1.2 + d_{pq}^{truck}$. Hence $\sum_{r \in R_{pq}} h_r = d_{pq}^{auto}$. Total link flows are the result of route flow aggregation, $f_a = \sum_{r \in R: a \subseteq r} h_r$. A solution is feasible if it respects the constraints on total origin flows, $\sum_{mq} d_{mpq} = d_{\bullet p \bullet}$, and on total destination flows, $\sum_{mp} d_{mpq} = d_{\bullet\bullet q}$, where $d_{\bullet p \bullet}$ and $d_{\bullet\bullet q}$ are given originating and terminating flows. MOD flows have the doubly-constrained logit form, $d_{mpq} = A_p \cdot B_q \cdot \exp(-\mu \cdot u_{mpq})$, where μ is a cost sensitivity parameter, and balancing factors, A_p, B_q, that ensure the constraints hold on total origin and destination flows. Therefore, OD flows are a decreasing function of generalized travel cost; the exponential cost function may be expanded to include a nested logit function for mode choice; the multinomial function is assumed here. These functions may be motivated in various ways, including random utility theory and entropy-based methods (Erlander and Stewart, 1990).

Model inputs (u_{mpq}) refer to transit and road levels of service. Transit costs are in-vehicle travel time c_{tpq}^{ivtt}, out-of-vehicle travel time c_{tpq}^{ovtt}, and fare c_{tpq}^{fare}, for travelling from origin p to destination q by transit, which are fixed regardless of flow. Origin-destination generalized cost by transit

u_{tpq} is a weighted sum of the three components plus a constant bias. Auto travel time on link a is an increasing function of total link flow, $tt_a(f_a) = tt_a^0 \cdot (1 + 0.15 \cdot (f_a/k_a)^4)$, where tt_a^0 and k_a are respectively the free-flow travel time and capacity of the link. Link generalized cost is $t_a = tt_a(f_a) + 0.15 \cdot l_a$, where l_a is the link length and the coefficient, 0.15, reflects a combination of both the direct effect of distance on generalized cost and the indirect effect of fuel consumption. Fixed additional auto costs, ac_{apq}, account for the parking fees and out-of-vehicle travel times at the origin and destination. The route generalized cost is $c_r = ac_{apq} + \sum_{a \subseteq r} t_a$. The minimum OD generalized cost by auto is $u_{apq} = \min\{c_r : r \in R_{pq}\}$. For every route $r \in R_{pq}$, define the *excess cost* as: $ec_r = c_r - u_{apq}$. The user-equilibrium assumption is that the excess cost of every used route is zero.

3. Solution Properties of Static Models for the Chicago Regional Network

To provide a general background for interpreting the findings for the route pattern, we present selected overall measures of the model solution for travel by auto, transit and truck in Table 1 for three values of the cost sensitivity parameter μ. The values were chosen so that the cost sensitivity doubles from the first to second to third solutions. The table gives the total person flows by mode and vehicle flows by auto and truck between zone pairs (interzonal) and within zones (intrazonal), and the total generalized costs for the same categories including parking fees and out-of-vehicle time. The mean values shown are the ratio of total costs to total flows.

The lowest cost sensitivity (CS) value (0.05) has the highest travel cost and congestion level, since travelers are relatively insensitive to travel costs. As a result, the percentage of travelers using transit is highest for this value, and transit is more attractive to higher cost trips. Therefore, mean transit costs rise together with mean auto costs. The lower half of the table showing intrazonal trips is less important, since the cost of auto travel within zones is fixed. Nevertheless, some insights may be gained by studying these values in association with the upper half of the table.

Table 2 shows the number of OD pairs and total number of routes in the three solutions grouped by the number of user-equilibrium routes per OD pair. These groupings are defined by order of magnitude, except for the first four rows, which are shown in more detail. Notice how the number of OD pairs with one route increases from 0.9 M to 1.8 M as the cost sensitivity increases from 0.05 to 0.20. The number of OD pairs with only one route may be regarded as surprising and quite remarkable. Likewise, it is completely unexpected that

Table 1. Summary Measures for the Chicago Region for the Morning Peak Period, 6:30–8:30 am

Zone Pairs	Cost Sensitivity	Persons			Vehicles	
Interzonal		auto	transit	transit (%)	auto	truck
Flow	0.05	903,191	590,691	39.5	752,659	418,507
(__/hour)	0.10	1,021,740	456,588	30.9	851,450	418,507
	0.20	1,116,689	331,028	22.9	930,574	418,507
Total cost	0.05	623,863	434,891	41.1	519,886	244,357
(hours)	0.10	473,016	256,196	35.1	394,180	224,603
	0.20	363,314	147,549	28.9	302,762	210,387
Mean cost	0.05	41.4	44.2		41.4	35.0
(minutes)	0.10	27.8	33.7		27.8	32.2
	0.20	19.5	26.7		19.5	30.2
Intrazonal		auto	transit	transit (%)	auto	truck
Flow	0.05	18,249	1,081	5.6	15,207	26,678
(__/hour)	0.10	33,887	996	2.9	28,239	26,678
	0.20	64,971	523	0.8	54,142	26,678
Total cost	0.05	5,145	612	10.6	4,288	6,116
(hours)	0.10	8,446	486	5.4	7,039	6,116
	0.20	13,895	214	1.5	11,580	6,116
Mean cost	0.05	16.9	34.0		16.9	13.8
(minutes)	0.10	15.0	29.3		15.0	13.8
	0.20	12.8	24.5		12.8	13.8

147 OD pairs have more than 100,000 routes each, comprising 13% of all routes in the network. The maximum number of routes per OD pair for the three solutions is: 459,264 routes for 0.05 (four OD pairs with origin 1777); 21,360 routes for 0.10; and 1,920 routes for 0.20. The location of these three cases is shown in Figure 1. See also Figure 2 for a map of the road network.

4. Solution Properties of Traffic Equilibrium Models

In this section, two sets of charts illustrating the solution properties of the model described above are presented. The first set of charts, Figures 3–10, show the Number of OD Pairs in the upper figure, and the Total Number of Routes in the lower figure, for a specified Number of Routes per OD Pair. Also shown on these figures is the Cumulative Proportion of OD Pairs and Total Routes. Figures 3 and 4 pertain to cost sensitivity of 0.05, the lowest value considered to date, which has the largest number of routes. Figures 5 and 6 show comparable

Table 2. Number of OD Pairs and Total Routes by Cost Sensitivity and Routes per OD Pair

	Cost Sensitivity			Cost Sensitivity		
	0.05	0.10	0.20	0.05	0.10	0.20
Routes per OD pair	Number of OD pairs			Number of routes		
1	897,845	1,366,377	1,781,131	897,845	1,366,377	1,781,131
2	573,978	747,165	724,829	1,147,956	1,494,330	1,449,658
3 to 5	599,977	533,131	412,279	2,202,263	1,983,654	1,517,572
6 to 9	364,524	262,007	150,810	2,554,880	1,838,277	1,023,664
10 to 99	623,431	271,100	131,306	17,747,585	6,415,025	2,460,034
100 to 999	122,424	21,314	1,950	35,746,073	5,055,657	362,711
1,000 to 9,999	17,789	1,210	5	47,537,050	2,375,019	7,256
10,000 to 99,999	2,195	6	0	58,036,340	90,660	0
100,000 to 999,999	147	0	0	25,032,733	0	0
Total	3,202,310	3,202,310	3,202,310	190,902,725	20,618,999	8,602,026
Routes per OD pair	Percentage of OD pairs			Percentage of routes		
1	28.0	42.7	55.6	0.5	6.6	20.7
2	17.9	23.3	22.6	0.6	7.2	16.9
3 to 5	18.7	16.6	12.9	1.2	9.6	17.6
6 to 9	11.4	8.2	4.7	1.3	8.9	11.9
10 to 99	19.5	8.5	4.1	9.3	31.1	28.6
100 to 999	3.8	0.7	0.1	18.7	24.5	4.2
1,000 to 9,999	0.6	0.0	0.0	24.9	11.5	0.1
10,000 to 99,999	0.1	0.0	0.0	30.4	0.4	0.0
100,000 to 999,999	0.0	0.0	0.0	13.1	0.0	0.0
Total	100.0	100.0	100.0	100.0	100.0	100.0

results for a cost sensitivity of 0.10, and Figures 7 and 8 show results for a cost sensitivity of 0.20. Figures 9 and 10 present a composite result for these three cost sensitivity values. A discussion of these results follows these eight figures.

Let us first consider Figures 3 and 4 as examples. In the upper left-hand corner of Figure 3, we see a point representing OD pairs connected by one route; as indicated in Table 2, 897,845 OD pairs of a total of 3,202,310 interzonal pairs are connected by one route. The next point to the right shows there are 573,978 OD pairs with two routes. Continuing in this manner to the lower right-hand corner, we note there are four OD pairs with the maximum number of routes, 459,264. Moreover, there are many instances of only one OD pair with a certain number of routes, as shown by the numerous points on the horizontal axis. Above that row of points, we can distinguish sets of OD

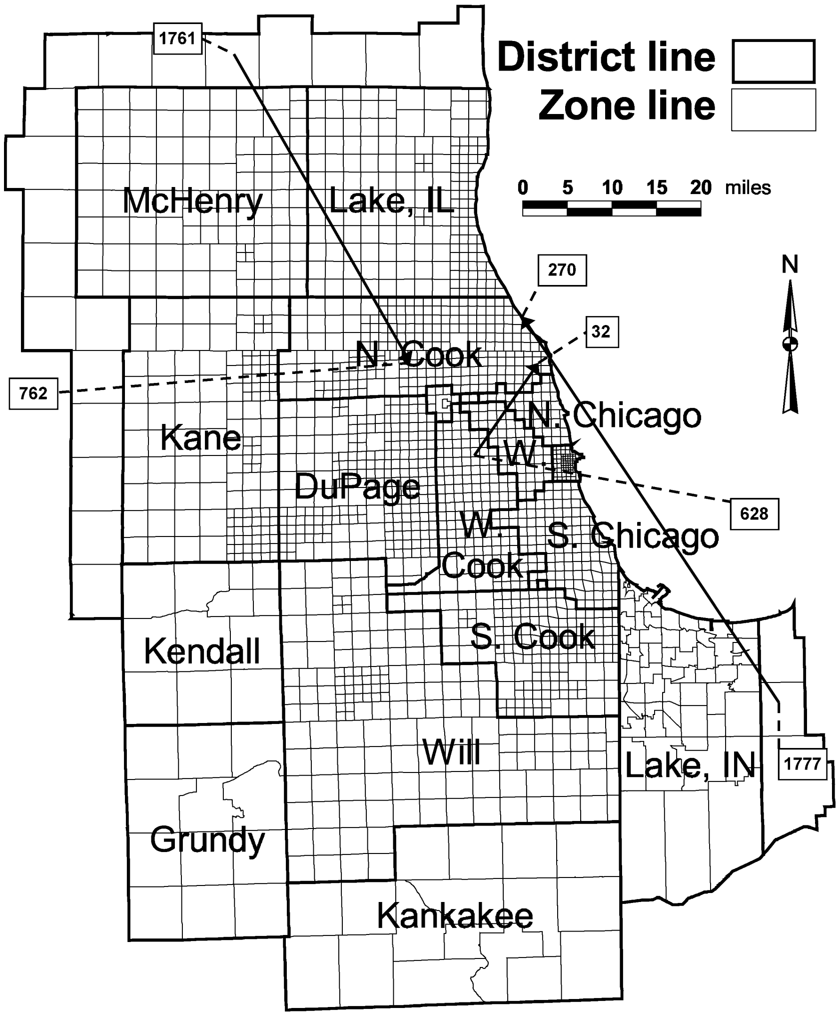


Figure 1. Zone System of the Chicago Region

pairs with two, three, four and five routes. Note that the use of the log-log scale to present these results tends to downplay the enormous numbers of routes and OD pairs; however, this scale appears to be the only practical way to present all the data in one figure. Also shown in Figure 3 is the Cumulative Proportion of OD Pairs plotted on the right-hand vertical axis in reverse order. The four OD pairs with the maximum number of routes are shown in the lower right as a proportion of 1.25E-06, and the OD pairs with only one route make up the last proportion of 0.280 in the upper left.

In Figure 4, we see the same results presented in a way that emphasizes the number of routes. As before, OD pairs with one route appear in the upper

Figure 2. Roadway Network of the Chicago Region

left-hand corner as point (897,845, 1). But now the second point appears as (1,194,756, 2), since the total number of routes with two routes per OD pair is twice the number of OD pairs. The case with the smallest number of total routes is a single OD pair with 223 routes, shown as the lowest point in the figure. The case with the largest number of total routes, 3,906,136, is 71 OD pairs with 55,016 routes each, which can be seen in Figure 3 as the point very close to the plot of the Cumulative Proportion of OD pairs. In Figure 4, the Cumulative Proportion of Total Routes is also plotted in reverse order.

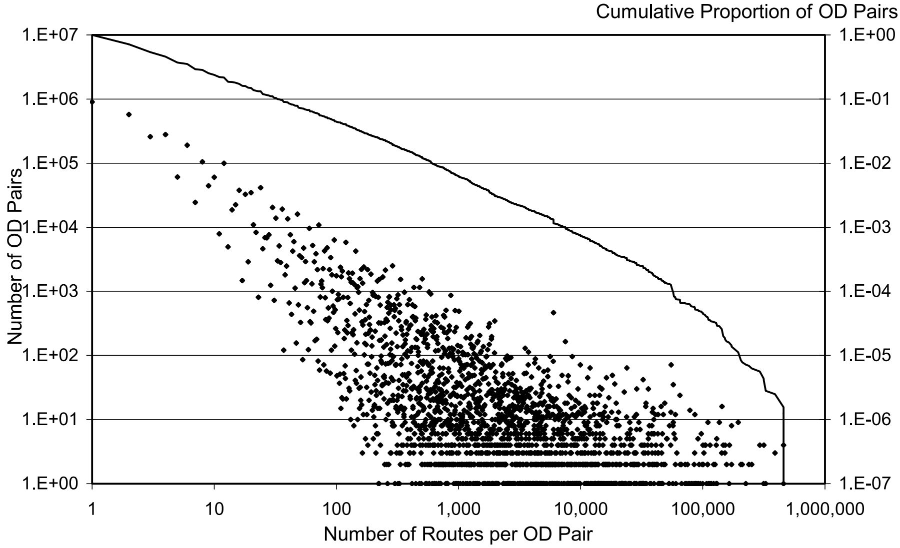


Figure 3. Number of OD Pairs by Number of Routes for CS - 0.05

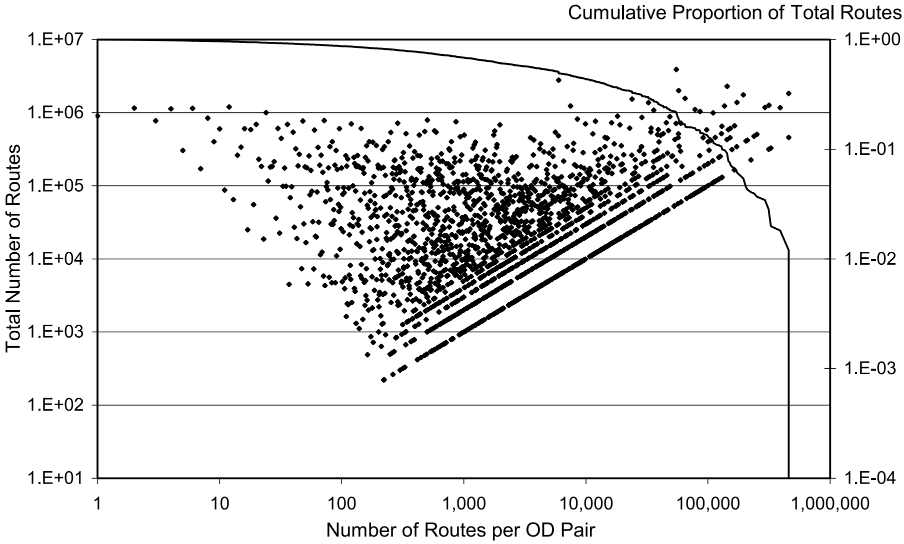


Figure 4. Total Number of Routes by Number of Routes for CS - 0.05

The corresponding figures for cost sensitivity values of 0.10 and 0.20 permit us to compare the Number of OD Pairs and Total Number of Routes for these three solutions. Note that all charts are plotted on the same axes to permit easy comparison. The number of OD pairs with only one route increases

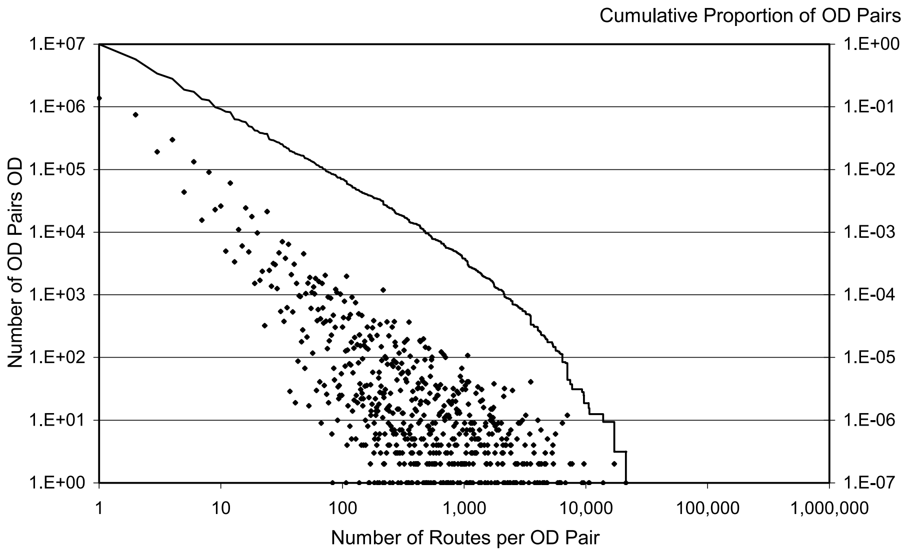


Figure 5. Number of OD Pairs by Number of Routes for CS - 0.10

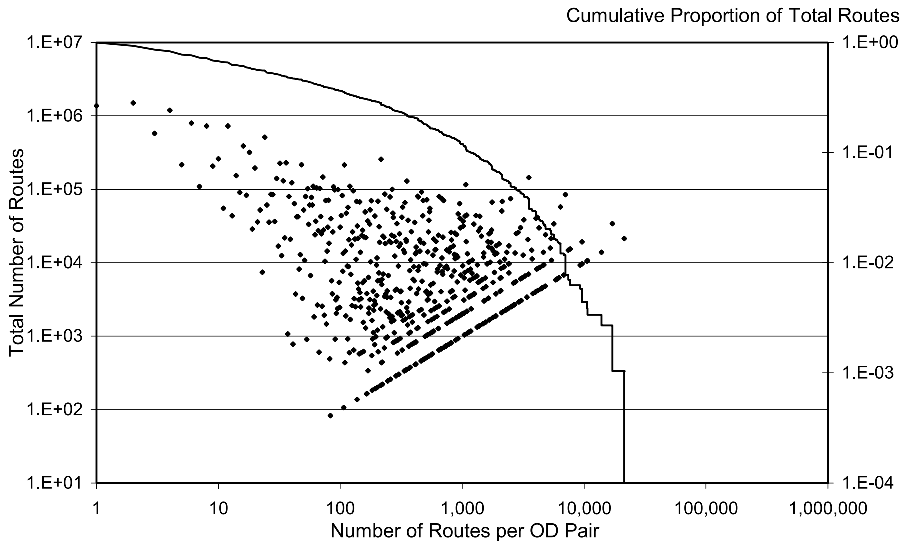


Figure 6. Total Number of Routes by Number of Routes for CS - 0.10

substantially as the cost sensitivity increases, as was also shown in Table 2. Likewise, the maximum number of routes decreases from 459,264 to 21,360 for CS of 0.10 to 1,920 for CS of 0.20. Finally, Figures 9 and 10 attempt to

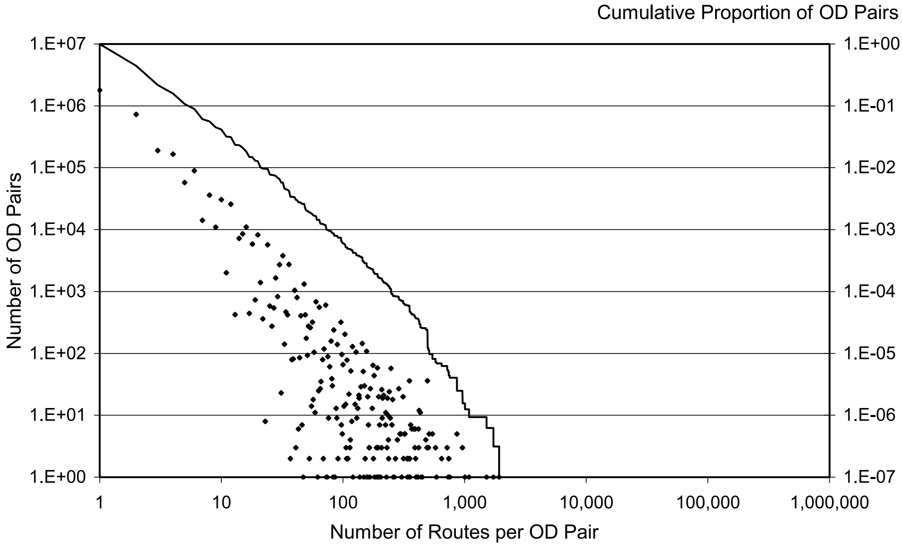


Figure 7. Number of OD Pairs by Number of Routes for CS - 0.20

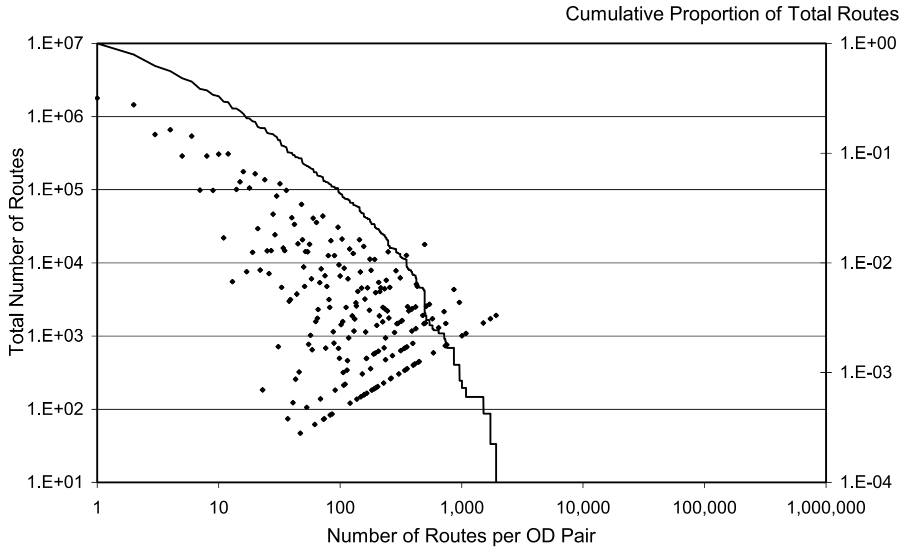


Figure 8. Total Number of Routes by Number of Routes for CS - 0.20

portray all three solutions on the same charts for easier comparison; however, some points are obscured by overprinting.

These initial results suggested several questions for further study. The first, which is explored here, concerns the generalized travel costs corresponding

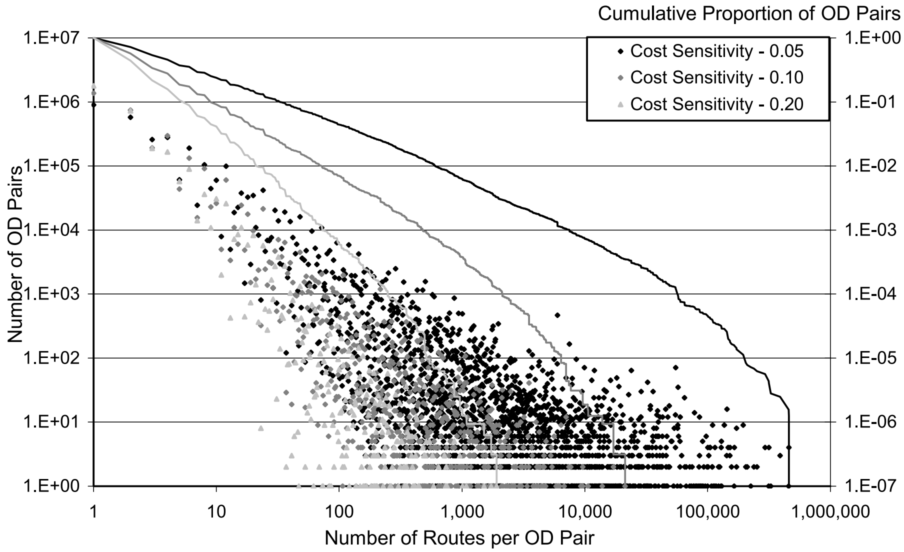


Figure 9. Number of OD Pairs by Number of Routes per OD Pair

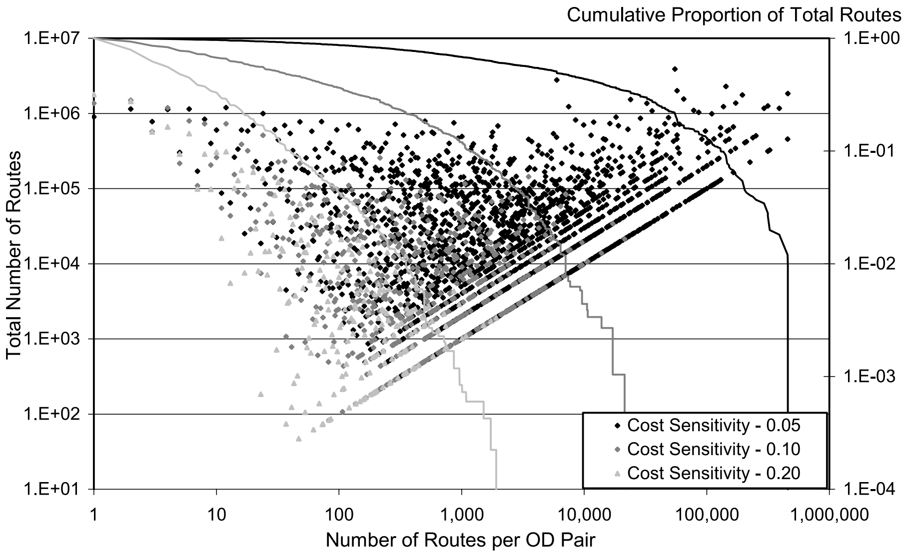


Figure 10. Total Number of Routes by Number of Routes per OD Pair

to these OD pairs. Are the unexpectedly large number of OD pairs with only one route simply connecting nearby OD pairs? Or, are there actually widely separated OD pairs connected by only one route in the traffic network equilibrium? Likewise, what is the cost between those OD pairs with very large

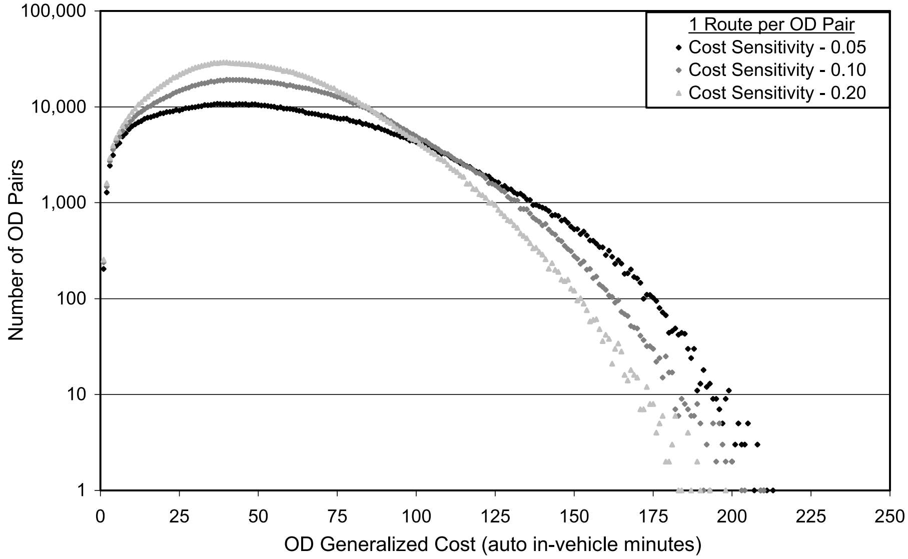


Figure 11. Number of OD Pairs by OD Cost: 1 Route per OD Pair

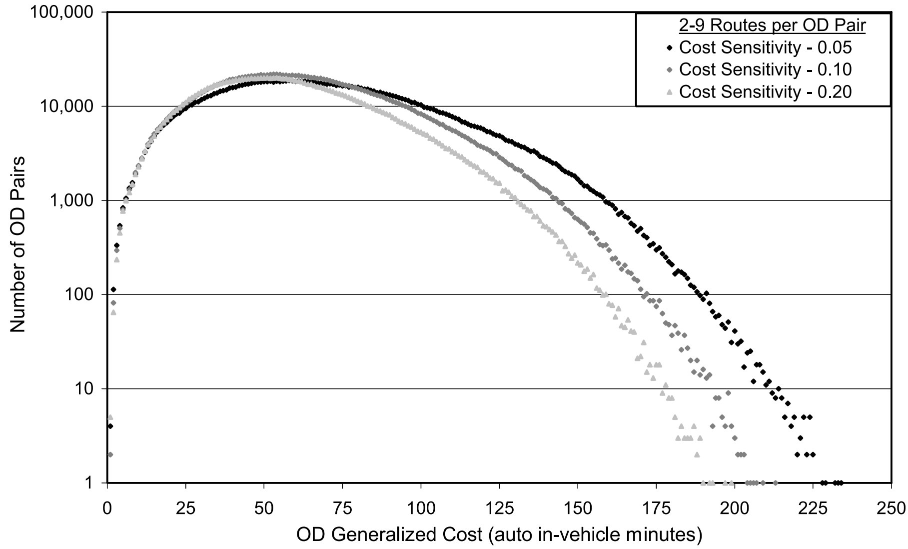


Figure 12. Number of OD Pairs by OD Cost: 2–9 Routes per OD Pair

number of routes? In the next set of charts, I explore the relation of the number of OD pairs to generalized travel cost for the three cost sensitivity solutions.

Figures 11–15 show the Number of OD Pairs by Generalized Travel Cost, stated in auto in-vehicle minutes, for each of five groupings of the Number of

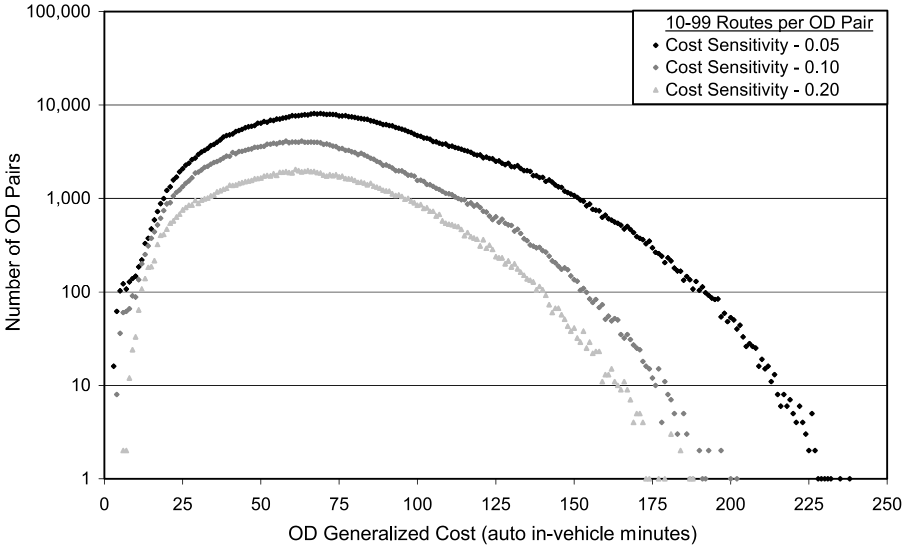


Figure 13. Number of OD Pairs by OD Cost: 10–99 Routes per OD Pair

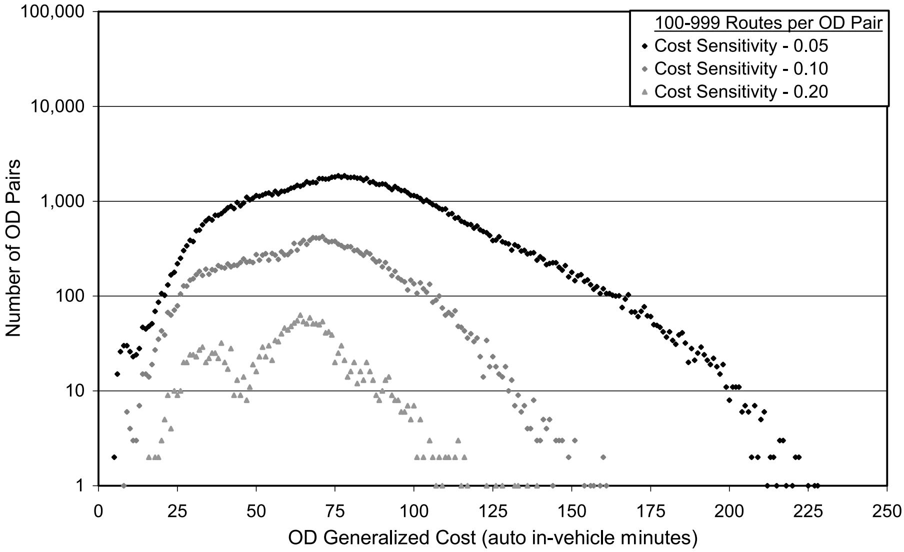


Figure 14. Number of OD Pairs by OD Cost: 100–999 Routes per OD Pair

Routes per OD Pair: 1 route; 2–9 routes; 10–99 routes; 100–999 routes; and 1,000 and more routes. These groupings were chosen following examination of finer groupings, and experimentation with various definitions. Figures 16–18 show the same results by cost sensitivity of the solution. All of the figures

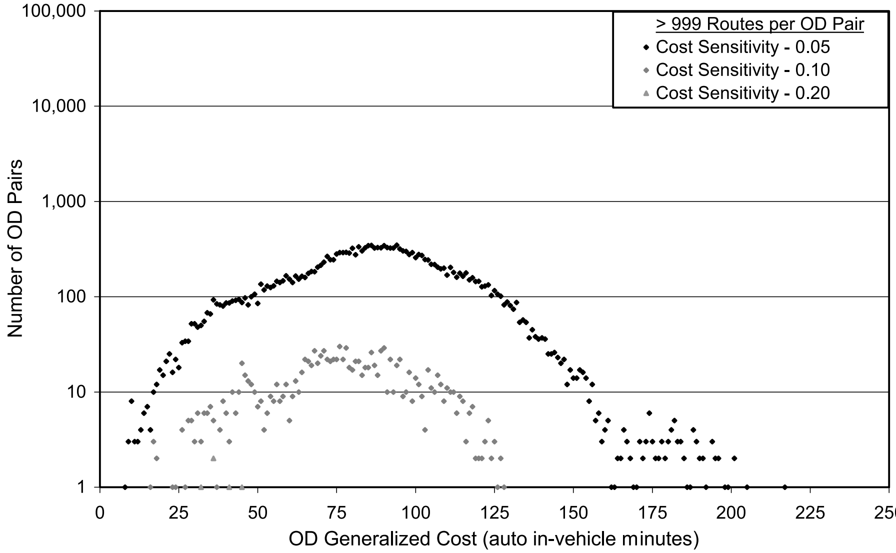


Figure 15. Number of OD Pairs by OD Cost: > 999 Routes per OD Pair

are plotted on the same scales of the horizontal and vertical axes for easy comparison.

Let us first consider Figure 11 showing the distribution of OD pairs with one route across the range of generalized costs, 0 to 240 minutes. The distribution of these OD pairs is remarkably broad, indicating at equilibrium that single routes connect OD pairs over a very wide range of costs, and definitely do not only connect nearby OD pairs. For the lowest value of cost sensitivity (0.05), the upper limit of cost reaches 213 minutes, whereas for the highest value (0.20), the upper limit is 198 minutes. The maximum number of OD pairs for the highest cost sensitivity of 0.20 reaches 29,000 pairs at 40 minutes, whereas for the lowest value, the maximum is about 10,700 pairs, also at 40 minutes. The intermediate cost sensitivity solution (0.10) lies between these two solutions.

For OD pairs with 2 to 9 routes, the three solutions effectively coincide from costs of 1 to 65 minutes. Then the solutions diverge with the lowest cost sensitivity solution extending to a maximum cost of 234 minutes. Figures 13–15 show a larger number of OD pairs and greater generalized cost for the 0.05 solution, as compared with the two other solutions. The shapes of the curves are similar to the groupings of 1 and 2–9 routes. One OD pair with more than 999 routes in the 0.05 solution has a cost of 214 minutes, as compared with a cost of only 34 minutes for the 0.20 solution. In studying the charts, the reader should recall that the higher costs in the 0.05 solution are the result of two interrelated factors: lower sensitivity to travel costs, and higher link and route costs resulting from increased traffic on the road network.

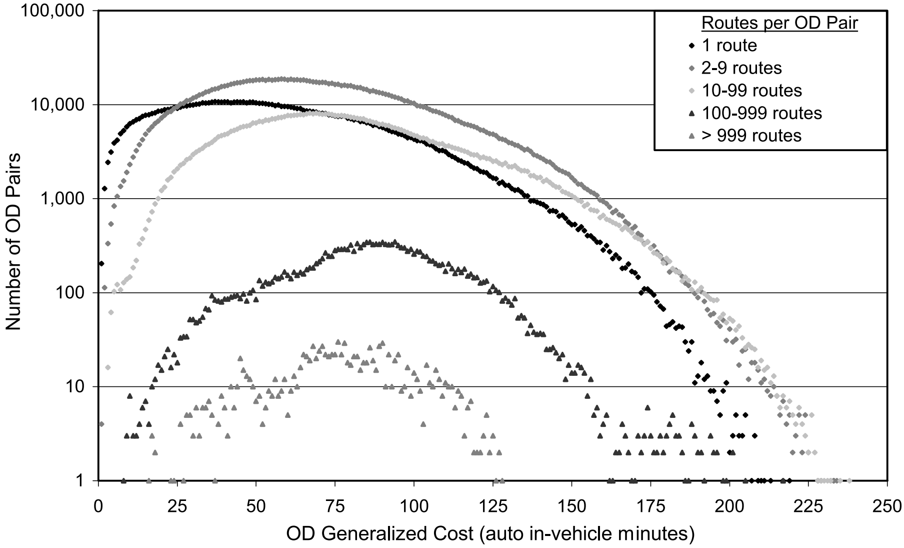


Figure 16. Number of OD Pairs by Cost and Routes: CS - 0.05

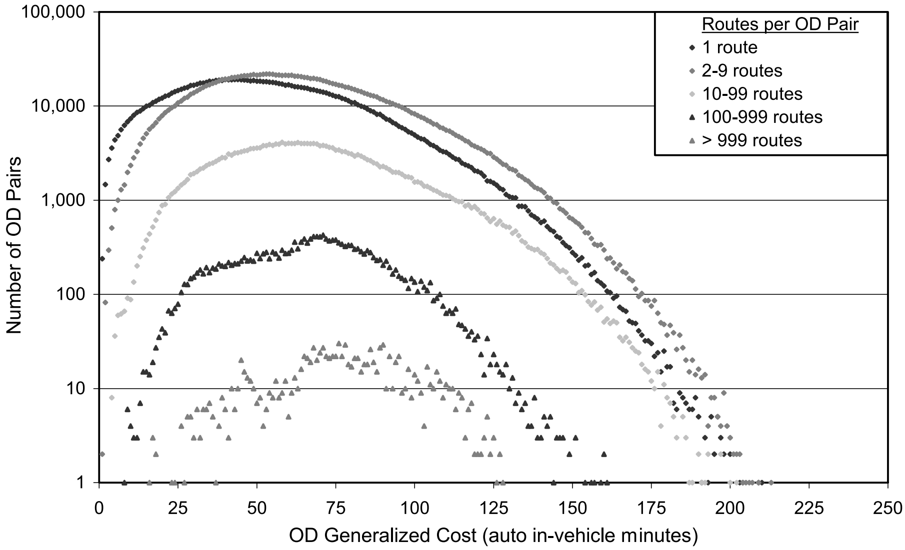


Figure 17. Number of OD Pairs by Cost and Routes: CS - 0.10

As noted above, Figures 16–18 reassemble the same results into one chart for each of the three solutions. By comparing these three charts, the reader can observe, for example, the number of OD pairs connected by one route is less dominant in the 0.05 solution than in the 0.20 solution. Accordingly, the

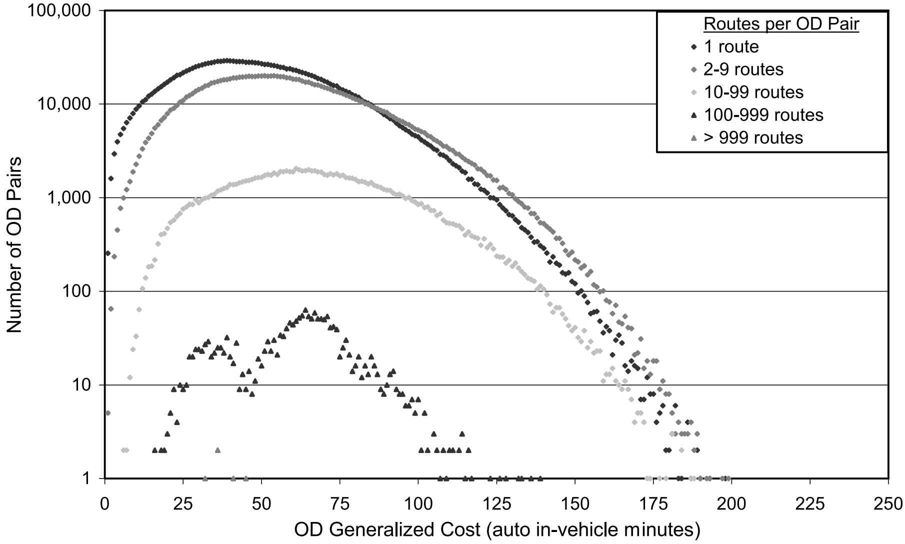


Figure 18. Number of OD Pairs by Cost and Routes: CS - 0.20

number of OD pairs with more than 999 routes is about 20,000 in the 0.05 solution, as compared with only five in the 0.20 solution. These charts should also be interpreted in conjunction with Table 2.

5. Patterns of Routes for OD Pairs with the Maximum Number of Routes

The maximum number of routes connecting an OD pair in the network is surprisingly large, and may appear to be implausible. To explore this property of the solutions, maps of the links used by any route connecting an OD pair were prepared. In Figures 19–21, the links corresponding to these routes are shown for three OD pairs: (1777–270), one of the four OD pairs with 459,264 routes, the maximum number in the solution with CS = 0.05; (1761–762), the OD pair with the maximum number of routes in a solution with CS = 0.14; (628–32), the OD pair with the maximum number of routes in the solution with CS = 0.20. The second case is included here because the OD pair for CS = 0.10 is very similar to that for CS = 0.05, and because the road network connecting (1761–762) is less grid-like than the other cases. See Figure 1 for the location of these OD pairs. Shown below are the cost sensitivity values, the number of routes, the user-equilibrium generalized cost of all routes, and the OD flow in vehicles per hour (vph).

Figure 19 shows the traffic equilibrium route structure for three model solutions for the zone pair which has the maximum number of routes in

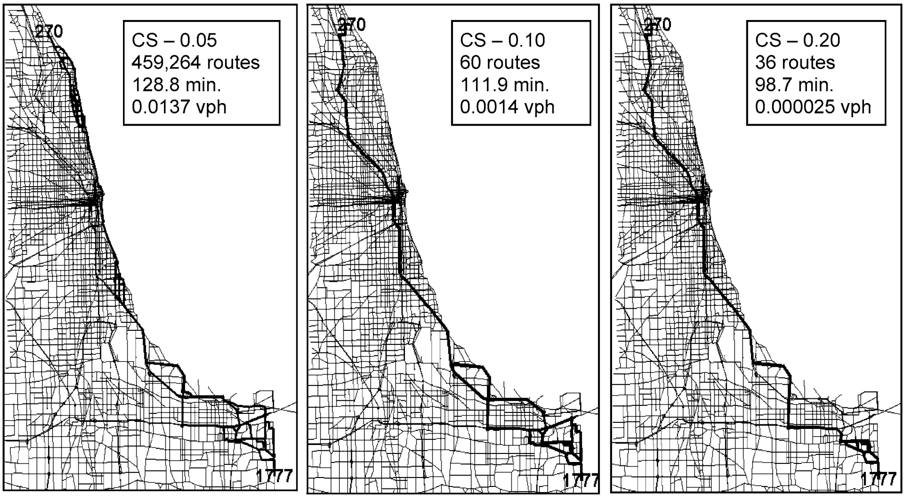


Figure 19. OD Pair 1777-270

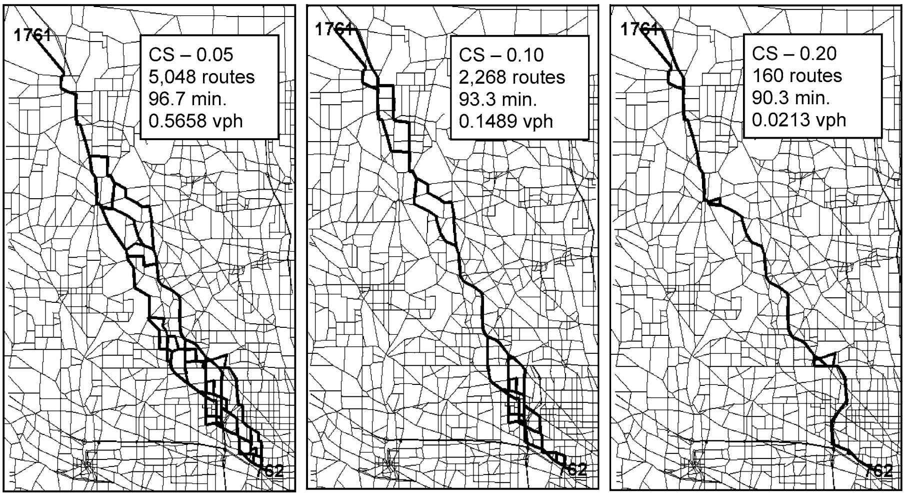


Figure 20. 20. OD Pair 1761-762

the solution with CS = 0.05. The left map shows the routes for the most congested solution, as compared with the two other solutions shown in the center and right. The routes shown in the right map are simple and highly plausible. The routes consist of a single sequence of links, except near the origin zone, where several alternative routes for traveling to I-94 occur. In the center figure, additional routes appear, again near the origin. Once the routes pass into Illinois at the southeast end of the long diagonal link, the

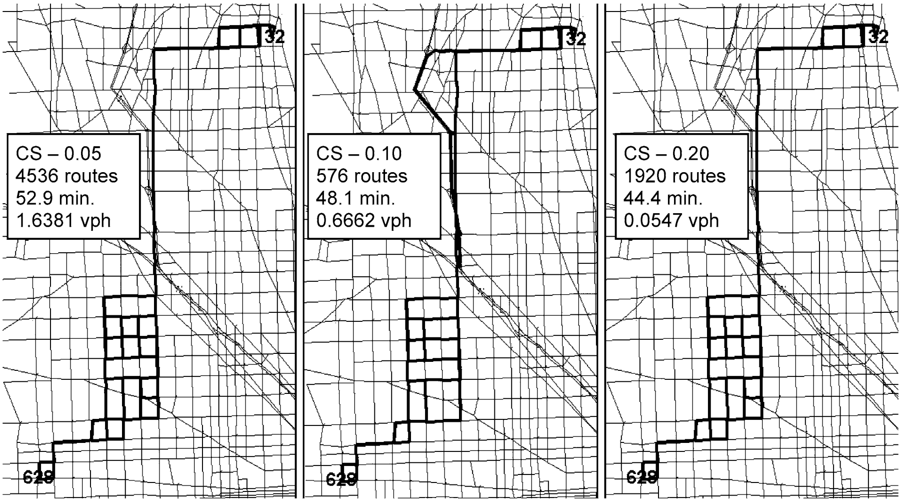


Figure 21. 21. OD Pair 628-32

Chicago Skyway, there is only one sequence of links. In the left map, a very rich route structure emerges in which the north-south expressway (I-90-94) is displaced by the Lake Shore Drive. Presumably this shift occurs because the expressways are heavily utilized by other flows in this rather congested solution. All of the route segments appear plausible, given the relatively simple network representation employed.

Figure 20 shows the route structures for the three solutions for a zone pair which had the maximum number of routes, 5,088, in another solution with CS = 0.14, not presented elsewhere in the paper. The right map is very simple, but includes a number of minor alternatives that result in 160 routes. The center map has additional alternative subroutes, which result in over 14 times as many routes as in the right map. The left map has a similar structure, but the total number of routes is 30 times the right map. All of the routes appear to be quite plausible.

Figure 21 shows the route structure for the zone pair with the maximum number of routes in the solution with CS = 0.20. This zone pair connects two inner suburbs, Oak Park to Evanston. The road network is a strong grid; many alternative routes with the same travel cost result from this structure. The principal difference between the right and center maps is the two alternative north-south routes near the destination; the right-hand route is an arterial road, and the left-hand route is the I-94 expressway. The absence of these two routes in the center map reduces the total number of routes by 70%. In the more congested left map, the number of routes is more than twice the right map, and the route structure has more major alternatives. However, the expressway route in the right map does not reappear, presumably because it is highly congested by travel between other OD pairs, just as in Figure 19. In the left map, many more grid street segments are used, which result in a large number of route combinations. Note that in each figure shown, the equilibrium travel costs increase from right to left by roughly equal intervals, and the flows also increase from right to left as a result of the decreasing cost sensitivity.

6. Conclusions from the Analysis

This exploratory research has extended the initial inquiry presented in Bar-Gera and Boyce (2005) by comparing three solutions to a traffic network equilibrium problem with variable flows. Only the cost sensitivity parameter values differ among the three solutions in order to study the effect of travel cost and congestion on the structure of equilibrium routes. As the number of solutions studied expands, the characteristics are consistent and predictable. However, the number of routes in the network equilibrium is large and perhaps unexpected.

A natural question about the results concerns which solution most closely represents reality in the Chicago region. Based on estimation of the parameters from 1990 data (Boyce and Bar-Gera, 2003), we would suggest the cost sensitivity parameter value most likely lies between 0.1 and 0.2. This statement, however, is a substantial oversimplification of the complexity of models actually applied in travel forecasting. Those models recognize that travel for each purpose and by each socio-economic group has its typical cost sensitivity, and that it is inappropriate to group these travelers into a single class. Investigation of a multi-class model, then, is one next step that ought to be pursued.

Another research question concerns the effect of the Chicago road network, and the network coding procedure, on the findings. Two experienced practitioners have suggested that the results presented here may be highly dependent on such attributes. Analysis of more road networks of various layouts is therefore indicated. A related question concerns the treatment of intersections and turning movements in the solution. In the present analysis no turning delays were assumed. Practitioner models include turning penalties and turn prohibitions, in order to reproduce observed link flows. We prefer to utilize turn-specific cost functions, but these have not been introduced so far. Therefore, much detailed investigation awaits us in the next round of analysis.

7. Implications for Other Types of Networks

In the context of this book on Network Science, it is appropriate to explore and speculate on whether and how the findings for traffic network equilibrium presented above apply to other types of networks. To this end, we offer a brief review of the literature of which we are aware. The idea of examining networks from a broad perspective is represented in the writings of several authors. Nagurney and Dong (2002) synthesized many network-related problems in their recent book, *Supernetworks*. Barabási (2002) also sought to relate a variety of networks to the experience of individuals through their personal experiences with urban traffic, air travel and the Internet. As a point of departure, we list several types of networks that share some characteristics with the type of network studied in this paper:

1. urban traffic networks
   - personal travel (personal auto; public transit; taxi)
   - trucking (parcel delivery; motor carrier; private trucking)
2. interregional transportation networks
   - personal travel (airlines and general aviation; auto; bus; rail passenger; ship)
   - freight shipments (airline; parcel delivery; motor carrier; pipeline; private trucking; railroad; barge and ship, and intermodal combinations of these *modes*)
3. supply chain management networks (management of the physical flows of goods, and the virtual flow of information, between and among stages in a supply chain)
4. utility networks
   - water
   - natural gas
   - electricity
   - district heating
5. telecommunication networks
   - telephone
   - data networks
   - cable

In considering these networks, one key distinction concerns whether the shipments are origin-destination-specific, or whether the shipments are interchangeable or *fungible*. Early problems in transportation, such as the classical Transportation Problem of Linear Programming and the Spatial Price Equilibrium Problem, assumed that shipments were interchangeable. Such shipments are termed *single commodity*. In contrast, most actual transportation problems encountered in reality are not interchangeable, and therefore are termed *multiple commodity*.

Utility networks are nearly always single-commodity networks, although there may be some distinctions about natural gas shipments that are unknown to us. Freight transportation shipments, in contrast, are generally not interchangeable, although grain shipments that comply with quality thresholds are fungible. Moreover, telecommunication networks, which are sometimes classified as utilities, are certainly not engaged in transmitting interchangeable commodities.

A second distinction concerns the agents who control and use the networks. Each type of network tends to have a distinctive terminology for its agents, so generalizations are problematic. Adopting transportation terminology, nevertheless, may get us started in classifying agents. The operator of the transportation network is generally called the *carrier*. The carrier provides the services, and sets the price in terms of the shipment characteristics (weight, distance, transit time, frequency of service, loss and damage, reliability, etc.), subject to possible *regulation* by a public body. The user of the services is called the *shipper*, which determines the amount to be shipped, the timing of the release of the shipment to the carrier, the choice of the type of carrier to be used, and other details such as packing specifications and insurance to reduce risk of loss and damage.

In some cases, such as personal travel by automobile, the shipper and carrier may be the same entity; in such cases, the carrier uses an infrastructure system by paying a toll or a user fee. Similar examples may be found for other networks.

Drawing on the experience with transportation networks, the behavior of shippers and carriers may be represented as constrained optimization problems, or generalizations of these, such as variational inequalities and fixed point problems. To formulate the problem, one must decide whether a hierarchy exists among the agents, as in a hierarchical game, or alternatively whether they interact on the same level. For example in freight transportation networks, one can define the following hierarchy:

Regulation:	international, national and regional regulatory agencies (world and continental trade organizations; national regulatory agencies; state/provincial regulatory agencies)
Carriers	modal and intermodal corporations competing for freight traffic
Shippers	corporations, associations and private individuals requiring freight services

Each of these levels can be represented by an optimization problem, for example, with sets of variable controlled by each level. Alternately, if the agents' interactions are all on the same level, their optimality and equilibrium conditions can be aggregated to form a single set. For example if the conditions of each agent are represented by a variational inequality, they can be summed to form one overall expression. We next examine some examples of problem formulations from transportation, supply chain, utility and telecommunication data networks.

Urban Traffic and Intercity Travel Networks

Only a brief indication of the large body of accumulated research findings on this topic will be mentioned here. The original model formulation achieved by Martin Beckmann (Beckmann et al., 1956) may be regarded as the genesis of the entire field of network models that combines the behavior to travelers and other decision makers with the passive operation of a transportation network. See Boyce (2007) for an account of the history of this development, and Boyce and Bar-Gera (2004) for a review of multiclass models. Unlike classical network models in which travel costs were regarded as fixed, Beckmann et al. succeeded to formulate two models with flow-dependent cost functions, and to show that their solutions correspond to: (a) the behavioral hypothesis that all used routes have equal and minimal travel costs, which is now called *user-equilibrium* or *user-optimal* behavior; (b) the minimization of total travel cost, which is now called *system-optimal*. Their models were also relatively general in that origin-destination flows (demand) were functions of travel costs, rather than being exogenous.

Beckmann and his coauthors investigated these conditions in terms of two model formulations, derived their optimality conditions, and showed that the system-optimal solution could be achieved in a user-optimal road network by imposing a set of tolls. Earlier, Wardrop (1952) had described these two behavioral hypotheses, but unlike Beckmann he proposed no model formulation; however, Wardrop's name is now generally associated with these concepts.

An extensive synthesis of urban traffic network models is provided by Patriksson (1994). For a more recent review, see Florian and Hearn (1999). Marcotte and Nguyen (1999) also provide a collection of papers that review this area.

Freight Transportation Networks

Several notable attempts to formulate and solve freight transportation models may be found in the literature beginning in the 1960s. Perhaps the first successful effort to represent the behavior of shippers and carriers in a single model system, and to implement and solve the formulation empirically was by Friesz and his collaborators in the early 1980s. Friesz et al. (1986) is representative of this research. Crainic (1999) provides an extensive review of this area, including his own extensive contributions.

The authors formulated two models, one representing shipper behavior, and the second representing the behavior of carriers. In the former model, shippers seek to minimize the delivered price of their shipments, as well as determine the shipment pattern of commodities from origins to destinations.

Price is defined as a weighted sum of transit time and a portion of the monetary cost of shipment charged by the carrier. Since shippers compete for possibly scarce freight transportation capacity, their choices of mode and carrier are *user-optimal* with regard to their shipper's network. The model is similar to a multi-class version of the model studied earlier in this paper for personal travel.

Carriers, in contrast, are represented as routing shipments over more detailed networks so as to minimize their total operating costs. In this sense, their routing decisions are *system-optimal*. The authors clearly defined the distinction of user-optimal and system-optimal in a system of freight transportation models, and succeeded to implement and solve the model for 15 groups of commodities over the U.S. railway network.

Supply Chain Management Networks

Network models that aim to manage the flow of physical products and associated information have come of age only recently, although early papers in this field date from the early 1970s; see the annotated bibliography of Geunes and Pardalos (2003). Supply chain networks seek to depict the relationships among suppliers, assemblers of products, wholesalers, retailers and finally consumers. The advent of the Internet had a profound effect on all stages of this process, and the transfer of supply chain functions to Internet websites may be expected to continue indefinitely into the future.

Nagurney has recently expounded on questions related to the conceptual and mathematical representation and solution of problems on such networks (Nagurney et al., 2002; Nagurney, 2005). Friesz et al. (2004) have also been active in exploring the mathematical foundations of this area of research.

Utility Networks

Electrical power systems offer another fascinating opportunity to explore and apply large-scale network models with a variety of economic interpretations. Patriksson (1994, pp. 56–57) offers a brief review and 20 references, noting that Duffin (1947) may have been the first to apply early notions of optimization to modeling electrical circuits. More recently, models of electrical power systems were proposed by Hobbs et al. (2000) and Hobbs (2001) based on Nash-Cournot game theory.

Nagurney and Matsypura (2006) describe a model of electric power generation, supply, transmission and consumption, drawing on their experience with supply chain network modeling. Their paper considers the behavior of several types of decision makers operating in a decentralized manner in a single variational inequalities formulation. The capability to represent the behavior of several agents in a rather general, consistent and interrelated way is impressive, and suggestive of applications to other types of network problems, such as the freight carrier-shipper problem already described.

The behavior of agents in their model with regard to profit maximization, pricing and production appears plausible. However, it would seem there must be alternative models with regard to price setters versus price takers, as well as production quotas, such as might be adopted by a cartel. These models would appear to include hierarchical relationships, which may be possible to represent in hierarchical variational inequalities formulations. This point also applies to supply chain models.

Data Networks

Bertsekas and Gallager (1987) synthesized an extensive range of material on data telecommunication networks, including their own journal papers; also, see Patriksson (1994, p. 58) for a short review, including eight references. Sansò and Soriano (1999) offer a recent collection of papers on the subject.

In Chapter 5, Routing in Data Networks, Bertsekas and Gallager formulated and investigated a system-optimal model of routing of data in congested telecommunication networks, which they note bears a strong resemblance to system-optimal transportation network problems; see, in particular, Section 5.5. Inasmuch as flows in data networks are very fast compared with transportation networks, the authors discuss with some care their assumption that the model considers average flows which change only very slowly over time, using the term *stationary* to describe this condition. Such models are often called *static* in transportation networks, as contrasted with more recent developments of *dynamic* models which are not considered here.

Two approaches for solving the routing model are described. The standard approach in practice at that time was to route data on the shortest time route. They state "A more sophisticated alternative is optimal routing based on flow models." This second approach is explored in terms of flow deviation and gradient projection methods. The first, which is standard in transportation network modeling, was also suggested for data networks by Fratta et al. (1973). The second algorithm, which has also been applied to transportation networks, was suggested for this problem by Bertsekas (1980).

Conclusions for Network Modeling

Finally, we reflect briefly on the implications of findings from our traffic equilibrium research for other areas of network modeling.

The first implication concerns whether the location and number of routes actually matter to the modeler and to the client for whom the model is solved. In the case of road traffic, the answer is certainly yes, routes do matter, especially if they are the most likely routes in terms of flow (Bar-Gera, 2006). A technique called Select Link Analysis is used by practitioners to determine which travelers benefit from a proposed network improvement, or which travelers would be impacted by a proposed road toll, for example. For the purpose of such analyses, the location of routes and the most likely route flows clearly do matter.

In freight transportation network and supply chain network modeling, a similar interest in predicting which shipments are impacted by network links temporarily taken out of service because of derailments, floods, etc. is clearly important. Whether there are similar concerns for electric power networks is less clear, and depends on physical and economic considerations in power system operation and pricing that are beyond our expertise.

Second, the capability to solve network models to precise levels of convergence may well be important in other applications for reasons that are different from traffic networks. At this time there is no general agreement that precise solutions of traffic equilibria are useful. Until recently, however, the ability to solve these problems precisely did not exist, so practitioners were not able to investigate to what extent the observed differences in the performance of alternative plans were cause by the plans themselves, as contrasted with errors arising from the precision of the solution of their models. As these solution methods become available in practitioner software systems, it will be interesting to observe whether they result in changes to professional practice.

A third point concerns the application of models for planning system infrastructure versus operating the system. In road traffic networks, static models of network equilibria, similar to the one described in this paper, are used for long-range planning of infrastructure investment, and to some extent for shorter-range impact analyses. Dynamic models may be used in the future for real-time traffic operations, but to our knowledge are not used for this purpose presently.

Models of freight transportation networks and electric power systems are potentially useful for infrastructure planning. On the other hand, we have found no evidence that models of data networks were used in infrastructure planning for the Internet, which was largely a private and competitive undertaking. Likewise, we are unaware of models of airline system infrastructure being used by aviation agencies to plan regional, national or international airport systems. Differences in problem definition and decision making, then, are the keys to understanding why such models are useful in one field and not useful in another.

Acknowledgements

This paper reports on findings from an ongoing collaboration of the authors. The route maps were prepared by Mr. Sagi Zinger at the first author's university. We thank Ronald Eash and Dean Englund for comments on an earlier paper based on this research, and for encouragement offered by Martin Beckmann and Anna Nagurney.

References

Barabási, A.-L. (2002). *Linked*. Perseus Publishing, Cambridge, MA.

Bar-Gera H. (2002). Origin-based Algorithms for the Traffic Assignment Problem. *Transportation Science*, 36, 398–417.

Bar-Gera, H. (2005). Primal Method for Determining the Most Likely Route Flows in Large Road Networks. *Transportation Science* 40, 269–286.

Bar-Gera H. and D. Boyce (2003). Origin-based Algorithms for Combined Travel Forecasting Models. *Transportation Research* 37B, 405–422.

Bar-Gera, H. and D. Boyce (2005). User-Equilibrium Route Set Analysis of a Large Road Network. In: H. S. Mahmassani (ed), *Transportation and Traffic Theory: Flow, Dynamics and Human Interaction*, Elsevier, Oxford, pp. 673–692.

Bertsekas, D. (1980). A Class of Optimal Routing Algorithms for Communication Networks. *Proc. 5th International Conference on Computers and Communication*, Atlanta, pp. 71–76.

Bertsekas, D. and R. Gallager (1987). *Data Networks*. Prentice-Hall, Englewood Cliffs, NJ.

Boyce, D. and H. Bar-Gera (2003). Validation of Multiclass Urban Travel Forecasting Models Combining Origin-Destination, Mode, and Route Choices, *Journal of Regional Science*, 43, 517–540.

Boyce, D., and H. Bar-Gera (2004). Multiclass Combined Models for Urban Travel Forecasting, *Networks and Spatial Economics*, 4, 115–124.

Boyce, D. (2004). Forecasting Travel on Congested Urban Transportation Networks: Review and Prospects for Network Equilibrium Models. *Networks and Spatial Economics*, 7(2).

Beckmann, M., C. B. McGuire and C. B. Winsten (1956). *Studies in the Economics of Transportation*. Yale University Press, New Haven, CT.

Crainic, T. G. (1999). Long-haul Freight Transportation. In: R. W. Hall (ed), *Handbook of Transportation Science*, Kluwer, Norwell, MA, pp. 433–491.

Duffin, R. J. (1947). Nonlinear Networks IIa. *Bulletin of the American Mathematical Society*, 53, 963–971.

Florian, M. and D. Hearn (1999). Network Equilibrium and Pricing. In: R. W. Hall (ed), *Handbook of Transportation Science*, Kluwer, Norwell, MA, pp. 361–393.

Fratta, L., M. Gerla and L. Kleinrock (1973). The Flow Deviation Method: An Approach to Store-and-Forward Communication Network Design. *Networks*, 3, 97–133.

Friesz, T. L., J. A. Gottfried and E. K. Morlok (1986). A Sequential Shipper-Carrier Network Model for Predicting Freight Flows. *Transportation Science*, 20, 80–91.

Friesz, T. L., R. Mookherjee, S. Peeta and P. Zhang (2004). Modeling the Dynamic Formulation of Distribution Networks and Supply Chains Using Self-Regulating Variational Inequalities. The Fifth Triennial Symposium on Transportation Analysis, Guadeloupe.

Geunes, J. and P. M. Pardalos (2003). Network Optimization in Supply Chain Management and Financial Engineering: An Annotated Bibliography. *Networks*, 42, 66–84.

Hobbs, B. F. (2001). Linear Complementarity Models of Nash-Cournot Competition in Bilateral and POOLCO Power Markets. *IEEE Transactions of Power Systems*, 16, 194–202.

Hobbs, B. F., C. B. Metzler and J.-S. Pang (2000). Strategic Gaming Analysis for Electric Power Systems: An MPEC Approach. *IEEE Transactions of Power Systems*, 15, 638–645.

Marcotte, P. and S. Nguyen (eds) (1998). *Equilibrium and Advanced Transportation Modelling*. Kluwer, Norwell, MA.

Nagurney, A. (1999). *Network Economics*, second and revised edition. Kluwer, Dordrecht, The Netherlands.

Nagurney, A. (2005). *Virtual Center for Supernetworks.* http://supernet.som.umass.edu/dart.html

Nagurney, A. and J. Dong (2002). *Supernetworks.* Edward Elgar, Northampton, MA.

Nagurney, A., J. Dong and D. Zhang (2002). A Supply Chain Network Equilibrium Model. *Transportation Research E*, 38, 281–303.

Nagurney, A. and D. Matsypura (2004). A Supply Chain Network Perspective for Electric Power Generation, Supply, Transmission and Consumption. In: E. J. Kontoghiorghes and C. Gatu (eds), *Optimisation, Econometric and Financial Analysis*, Springer, Berlin, pp. 3–27.

Patriksson, M. (1994). *The Traffic Assignment Problem: Models and Methods.* VSP, Utrecht.

Patriksson, M. and R. T. Rockafellar (2003). Variational Geometry and Equilibrium. In: *Equilibrium Problems and Variational Models*, Proceedings of the International School of Mathematics "G. Stampacchia" 32th Workshop, Kluwer Academic Publishers, Norwell, MA, pp. 347–368.

Sansò, B. and P. Soriano (eds) (1999). *Telecommunications Network Planning*, Kluwer, Norwell, MA.

Wardrop, J. G. (1952). Some Theoretical Aspects of Road Traffic Research. *Proceedings of the Institution of Civil Engineers*, *Part II*, 1, 325–378.

Appendix – Model Description

From H. Bar-Gera and D. Boyce (2005)

We consider a study area that is divided into a set of zones, Z, connected by transit services and by a road network. The road network consists of a set of nodes, N, and a set of directional links, A. In the Chicago regional model there are 1790 zones, 12,982 nodes and 39,018 links. A route is a sequence of nodes, $[v_1, \ldots, v_k]$, such that $[v_i, v_{i+1}] \in A$. The set of available routes from origin $p \in Z$ to destination $q \in Z$ is R_{pq}, and the set of all routes is R.

The purpose of the model is to predict: (a) the mode-origin-destination (MOD) flow d_{mpq}, in persons per hour, for every origin $p \in Z$, destination $q \in Z$, and mode $m \in$ {a=auto, t=transit}; (b) the distribution of auto OD flows to route flows h_r, for every route $r \in$R. Auto OD flows are the sum of person-trips divided by a constant auto occupancy factor (*aof* = 1.2) and the given OD truck flows d_{pq}^{truck}, in equivalent passenger cars per hour: $d_{pq}^{auto} = d_{apq}/aof + d_{pq}^{truck}$; hence $\sum_{r \in R_{pq}} h_r = d_{pq}^{auto}$. Total link flows are the result of route flow aggregation, $f_a = \sum_{r \in R: a \subseteq r} h_r$. A solution is feasible if it respects the constraints on total origin flows, $\sum_{mq} d_{mpq} = d_{\bullet p \bullet}$, and on total destination flows, $\sum_{mp} d_{mpq} = d_{\bullet\bullet q}$, where $d_{\bullet p \bullet}$ and $d_{\bullet\bullet q}$ are given inputs. In the model implemented for the Chicago Region, total MOD flows amount to 1.5 million (1,513,211) persons per hour, while total truck flows amount to 0.4 million

(445,184) passenger-car-equivalents per hour, representative of the morning peak period (6:30–8:30 am) in 1990.

The remaining model inputs refer to transit and road levels of service. Transit data are in-vehicle travel time c_{tpq}^{ivtt}, out-of-vehicle travel time c_{tpq}^{ovtt}, and fare c_{tpq}^{fare}, for travelling from origin $p \in Z$ to destination $q \in Z$ by transit. These are fixed regardless of flows. Origin-destination generalized cost by transit, u_{tpq}, is a weighted sum of the three components plus a constant bias; in the results reported here the transit bias is zero, and the weights are 0.25, 0.90, and 0.08 respectively.

Travel time by auto on link a, tt_a, is a function of total flow, $tt_a(f_a) = tt_a^0 \cdot (1 + 0.15 \cdot (f_a/k_a)^4)$, where tt_a^0 and k_a are respectively the free-flow travel time and capacity of the link. Link generalized cost is $t_a = tt_a(f_a) + 0.15 \cdot l_a$, where l_a is the link length and the coefficient, 0.15, reflects a combination of both the direct effect of distance on generalized cost and the indirect effect of fuel consumption. Fixed additional auto costs, ac_{apq}, account for the parking fee and out-of-vehicle travel time at the origin and destination. The route generalized cost is $c_r = ac_{apq} + \sum_{a \subseteq r} t_a$. The minimum OD generalized cost by auto is $u_{apq} = \min\{c_r : r \in R_{pq}\}$. For every route $r \in R_{pq}$, define the *excess cost* as: $ec_r = c_r - u_{apq}$. The user-equilibrium assumption is that the excess cost of every used route is zero. Approximate UE solutions are evaluated by the maximum excess cost over all used routes. MOD flows have the doubly-constrained logit form, $d_{mpq} = A_p \cdot B_q \cdot \exp(-\mu \cdot u_{mpq})$, with cost sensitivity $\mu = 0.2$, and balancing factors, A_p, B_q , that ensure the constraints hold on total origin and destination flows. Approximate solutions of this model are evaluated by total misplaced MOD flow, $\sum_{mpq} |d_{mpq} - A_p \cdot B_q \cdot \exp(-\mu \cdot u_{mpq})|$.

The combined model of user-equilibrium and mode-origin-destination choice with the specific structure described above can be formulated mathematically either as a fixed point problem, or as a convex optimization problem (Bar-Gera and Boyce, 2003). For the generalized link cost function stated above (separable, monotonically increasing), the equilibrium, or optimal, solution uniquely determines total link flows. Total link flows in turn uniquely determine link costs, route costs, and the set of minimum cost routes, referred to here as the set of UE routes. Of course, the route flows on this set of UE routes are not unique.

The model is solved by an origin-based assignment algorithm; route flow solutions are described by a set of restricting a-cyclic subnetworks A_p for each origin, and origin-based approach proportions $\alpha_{pa} \in [0, 1]$, such that $\alpha_{pa} = 0$ for all $a \notin A_p$, and $\sum_{a \in A_p : a_h = v} \alpha_{pa} = 1 \forall p \in Z, v \in N, v \neq p$. The implicit set of routes from origin p is the set of all routes within A_p, that is $R_{pq}[A_p] = \{r \in R_{pq} : a \subseteq r \Rightarrow a \in A_p\}$. The implicit route flows are given by $h_r = d_{apq} \cdot \prod_{a \subseteq r} \alpha_{pa}$. For the OBA algorithm, the number of UE routes per OD pair is unrelated to the number of iterations, unlike algorithms based on linearization methods.

The availability of route flows and approach proportions allows (a) adjustments of mode-origin-destination flows while retaining the current route proportions, (b) adjustments of restricting subnetworks to accommodate more routes, and (c) efficient adjustments to approach proportions that utilize second order derivatives of the objective function. The resulting algorithm offers precise convergence for large-scale networks. For the specific model of the Chicago Region presented here, the algorithm produced a solution with maximum excess cost of 1E-13 equivalent auto in-vehicle minutes of travel time and total misplaced MOD flow of 1E-10 person-trips per hour.

By itself, a precisely converged solution does not necessarily guarantee a set of routes that is similar to the true, unique set of UE routes. As link flows and link costs converge towards their equilibrium values, so should excess costs. Therefore, excess costs of UE routes should decrease continuously towards zero, while the excess cost of any non-UE route should converge to a strictly positive value. The minimum equilibrium excess cost of all non-UE routes, considered as the *rejection gap*, is strictly positive as well. In principle, if a threshold below the equilibrium rejection gap is chosen, then at a certain finite level of convergence, the excess cost of all UE routes will be below the threshold, and the excess cost of all non-UE routes will be above the threshold.

We chose to include all routes with excess cost below a threshold of 2E-12. The smallest excess cost of a rejected route is at least 30E-12, which determines the estimated rejection gap. Therefore, there are no routes with 2E-12 < excess cost < 30E-12. There are several reasons to believe that the chosen set of routes is probably similar and perhaps identical to the true set of UE routes. One reason is the stability of the set of included routes in the final iterations of the origin-based assignment. Another reason is the order of magnitude difference between the chosen threshold and the estimated rejection gap. Additional reasons are discussed in Bar-Gera (2006), particularly the fact that the chosen set of routes maintains consistent consideration of alternative route segments, a fundamental property of sets of minimum cost routes in general and the set of UE routes in particular.

Index

accidents, 8, 102, 123, 124, 305, 307
accuracy, 4, 80, 104, 107
acoustic emission, 107
actions, micro-level, 126
activity set, 61
adaptability, fast, 110
adaptation, 59, 118, 126
agencies, 196
  aviation, 331
  national regulatory, 327
  provincial regulatory, 327
  regional regulatory, 327
agent-by-agent, 79
agents, 4, 24–27, 53, 60, 77, 79, 83, 85–89, 91, 126, 127, 129, 187, 196, 327, 329, 330
  biological, 124
  classifying, 327
  competitive, 79
  disruptive, 196
  distributed, 126
  external, 65
  fundamental, 145
  game theoretic, 85
  infinite, 24
  intelligent, 99
  interacting, 126
  large number of, 124, 126
  route, 92
  selected, 86
  single, 27
agreement, 1, 95, 256, 331
AI algorithms, 73
airline, 18, 174, 224, 230, 232, 326
  tickets, 228, 232
  travel data, 16
airports, 16, 19, 20, 158, 159, 171, 188
Albert nodes, 174
algebraic
  relationships, 280
  simplifications, 277, 280
algorithm, 57, 72, 81–84, 94, 96, 112, 195, 201–203, 205, 207–214, 216, 217, 220, 270, 302, 306, 307, 330
  complementarity, 78
  decomposition, 73, 202
  development, 81
  diagonalization, 204
  efficient, 65
  genetic, 88
  nontraditional, 89
  searches, 208
  shortest-path, 169
algorithmic approach, effective, 73
alightings, 212, 220
alleles, 141, 146
allocations, 202, 228
Alonso model, 173
alpha parameter, 177
alternative estimation problem, 236
amplitude, 101
  high, 114
  low, 114
  vibration, 102
analogues, classical, 88
analysis
  comparative, 182
  comprehensive stability, 113
  control-theoretic, 110, 112, 118
  dynamic, 112
  empirical, 175
  fractal, 104
  frequency response, 112, 115
  functional, 86
  levels of, 168, 178
  local, 101
  numerical, 79
  qualitative, 79
  routing frequency, 181
  sensitivity, 244, 245
  shorter-range impact, 331
  social network, 158

structural, 175, 180, 181
anarchy, 26
price of, 24
application areas, 223, 227
applications, 3, 28, 49, 58, 72–74, 78, 79, 82, 83, 89, 92, 99, 104, 105, 107, 108, 114, 148, 153, 156, 158, 167, 269, 270, 287, 288
dynamic, 111
empirical, 147, 154, 169
engineering, 2, 100
medical, 102
network-based, 270
approaches, 11, 38, 39, 67, 73, 110, 111, 117, 168, 172, 182, 187, 194, 201, 203, 205, 208, 210, 220, 227, 228, 247–249, 330
algebraic, 117
alternative, 67, 68, 99, 188
closed-loop, 223, 227
computational, 92
fundamental, 79
holistic, 111
micro-simulations, 124
network-based, 202
range, potential, 188
scale-free, 183
standard, 230, 330
time-discretization, 291
approximate solutions, 334
approximation, 22, 42, 43, 65, 79, 188, 193, 287
finite difference, 88, 89
finite-difference, 73
formulas, 192
inner, 201, 205, 208
linear, 44
nonlinear function, 104
techniques, 193, 198
apriori
path generator, 72
reason, 89
arcs, 8–11, 15–17, 22–27, 62, 70, 72, 198, 205, 241
attacking, 196
bottom, 25, 26
capacities, 71, 197
number of, 9, 10
odd number of, 9–11
top, 25, 26
travel time functions, 241
ARMA model, 107
articulation, 54, 71
approximate mathematical, 80
aspect
game-theoretic, 223
socio-economic, 152
transient, 223
assignment, 134, 201, 205, 208, 209, 213, 214, 217
algorithm, 220
all-or-nothing, 209, 211, 213
problem, 203, 204
associations, 309, 327
assymptotic behavior, 35
asymmetric cases, 244
attacks, 20, 21, 59, 94, 157, 171
random, 21, 22, 157
targeted, 20–22, 157, 171
terrorist, 168, 189
attractiveness, 173
attractors, 32, 35–37, 44, 101, 106, 135
local, 36
multiple, 101, 103
pre- and post-bifurcation, 103
strange, 4, 38, 44, 45, 48, 58
term, 35
automobile, 201, 203, 220, 327
movements, 202
travel speeds, 219
travel times, 219
trip tables, 203
autonomous control strategies, 117
average
automobile travel speeds, 219
costs, 190
demands, 251
estimator, 195
minimum path, short, 147
prices, 251
width, 144
award, joint, 77

backlogs, 110, 112, 117
backorders, 90, 225
basis, 138, 154, 156, 169, 173, 175, 193, 194, 270
functions, 106
customized, 108
radial, 104

beer
- distribution game mimics, 110
- game approach, 111

behavior, 28, 30, 35, 45, 46, 50, 57, 58, 89, 99, 101, 103, 113–115, 126, 214, 220, 229, 260, 270, 273–275, 328–330
- aggregate, 58
- analysis, 113
- complex, 49, 58, 100, 112, 118
- dynamic, 103, 110, 112, 118
- ergodic, 129
- functional, 101
- game, 88
- irregular, 50, 100
- metastable, 134
- operational, 109
- optimizing, 83
- oscillatory, 114
- periodic, 50, 100
- shipper, 328
- similar, 83
- strange, 100
- strategic, 187, 196, 197
- traveler, 202
- types, 206, 215, 218
- unstable, 115
- user-optimal, 328

benefits, 69, 70, 110, 127, 174, 193, 224
- gross, 69

best response problem, 234, 240, 252
best-response demands, 238, 255, 262, 263
better fitness configuration, 138
bifurcation, 46–49, 58, 101, 112
- diagrams, 101,
- global, 101
- multiple, 103
- point, 101,
- saddle-node, 46–48
- transcritical, 47, 48

biological
- analog, 125
- terms, 146

biology, 79, 128, 141, 148
borderline, 44
- cases, 44

Boston subway network, 171
bottom tier, 271, 283
bounds, 26, 27
- stipulated, 93

Braess's paradox, 26, 27
branches, 80, 136–138, 203
breakage, 108
- chip, 102, 107

brevity, 81, 82, 84, 88
bridges, 8–10, 188, 196
broad-scale networks, 153
Brownian motion, 1, 96
budget line, 191
bullwhip effect, 111
bus, 204, 212, 214, 220, 326
- fares, 212, 216
- movements, 202, 204
- stops, 212, 214
- transfer time, 215
- vehicle loading, 212
- vehicles, 214, 220

business process reengineering, 111

cable, 326
- fiber optic, 70

calculations, 29, 302
- back-of-the-envelope, 93
- intermodal path, 220
- net benefit, 69
- static network equilibrium, 72

calculus, elementary, 7, 25
capacities, 23, 50, 68, 93, 112, 113, 124, 189, 190, 196, 214, 244–248, 255–257, 262, 264, 309, 334
- exceed, 237
- increasing traffic, 123
- limited, 53
- resource, 112
- scarce freight transportation, 329
- tight, 244, 245, 255, 257
- total, 229, 232

capacity
- change, 193
- constraint, 241, 255, 257
- perturbations, 71

capital budgeting, 60, 66, 67, 69, 70, 92, 93
- model, 69, 72
- problem, 94

carriers, 327–329
catastrophes, 58, 66
cause-and-effect relationships, 170
center, 35–39, 44, 123, 142, 323
- maps, 324
- traffic management, 124, 125

trajectories, 38
centralization, 168, 179
chain, 65, 70, 78, 80, 83, 86–89, 94, 201, 203, 207, 223, 239, 240, 269, 270, 273–278, 280, 282, 286, 289, 305, 327
example, 296
management, 110
network, 21, 182, 269, 271, 273, 276, 278, 283, 287, 291, 292, 296, 298, 299, 306, 329
equilibrium, 276, 278, 282, 283, 286, 287
model, 269, 271, 272, 276, 287
problems, 278, 290
examples, 269, 290, 291
flows, 293, 298
model formulae, 296
modeling, 329, 331
problems, 271
numerical examples, 271
strategies, 110
Champernowne deterrence function, 174
changes, large, 68
changing network conditions, 305, 307
chaos, 4, 28, 50, 51, 58, 99, 100
harnessing, 99
low-dimensional, 104–106
principles, 99
theory, 49, 50, 100, 105, 118
chaotic, 30, 44, 49, 50, 101, 104–106
attractors, 89, 101
dynamical system, 28
dynamics, 99
effects, 29
low-dimensional, 103
regimes, 109
characterization, 99–101, 103, 105, 107–109, 111, 113, 115, 117, 119, 121, 255
surface, 109
systems level, 99
charts, 310, 314, 316, 320–322
set of, 310, 318
chatter, 102, 103, 109, 112
suppress, 109
Chicago
region, 306, 307, 310, 312, 313, 325, 333
regional model, 333
regional network, 309
road network, 325
chief determinant, 102
choices, 27, 73, 78, 106, 107, 174, 193, 201, 202, 327, 329
mode-origin-destination, 334
transfer, 202
circles, 15, 21
social, 18
cities, 9, 16, 18, 57, 67, 152, 153, 171, 180, 181
peripheral, 171
civilian economy, 92–94
war-impacted, 93
classes, 14, 22, 28, 46, 48, 50, 65, 67, 77, 79, 92, 95, 153, 183, 271
single, 85, 325
classical noncooperative mathematical games, 84
closed form, 67, 94
solutions, 228, 252
clustering, 7, 13–15, 19, 20, 22, 154, 168, 169, 178
characters, 148, 151
close allowing, 15
coefficient, 15, 20, 22, 27, 147, 149, 151, 163, 178
local, 147
clusters, 159, 160, 170
coding scheme, simple, 131
coefficients, 104, 106, 107, 151, 155, 157, 229, 240, 244, 309, 334
power-law, 109
wavelet transform, 106
collection, 56, 57, 196, 328, 330
coupled, 56
collision avoidance system, 56
commercial Internet service providers, 94
commodities, 26, 60, 62, 63, 110, 328, 329
flows of, 61, 62
commuters, 123, 124, 126, 167, 174, 175, 177, 180, 182, 183
economic flows of, 180, 182
commuting, 123, 156, 167, 168, 173–175, 182
behaviour, 168
data, 180, 183
real, 177, 179
employees, 173
networks, 157, 168, 173, 174, 183
interpreting, 183
observed, 169, 176
real, 175
real-world, 179, 182
patterns, 177

complex, 168
home-to-work, 167
companies, 18, 223, 224, 232
comparative statics, 65
comparison, 67, 78, 117, 169, 177, 178, 183, 244, 314, 316, 320
first-cut, 182
supply-demand, 167
compensating variation, 191
competition, 54, 57, 58, 60, 77–81, 87, 92, 223, 225, 227, 228
biological, 77
collusive network, 66
dynamic, 54, 77
imperfect, 224, 225
competitors, 223, 224, 228–230, 232–234, 237, 238, 241, 245, 260
completeness, 292, 296
complex
adaptive system, 142
infrastructure systems, 53
network science approaches, 187
network theories, 167–169, 188
network-based problems, 269
networks, 1, 7, 16, 23, 24, 58, 59, 147, 148, 159, 167, 169, 170, 183, 187
large, 17, 170
systems, 46, 99, 148
system-wide behavior, 126
complexity, 1, 4, 46, 57, 99, 118, 195, 325
components, 9, 11, 20, 31, 56, 94, 102, 104, 170, 174, 187–189, 193–196, 240, 272, 295, 309
damaged, 93
surface-bearing, 109
undesired signal, 104
compound oscillations, 115
computable games, 53, 59
theory literature, 60
computation, 60, 77, 78, 92, 140, 191, 192, 233, 248, 270, 271, 290
multi-party, 58
computational
architectures, 58
difficulties, 83
economists solving, 78
impediment, 72
prowess, 4
results, 201, 205, 214, 220, 223, 227, 228, 243, 252, 260
schemes, 88
time, 217, 220
computer, 23, 29, 46, 95
military, 94
primitive, 307
science, 58
system, 291
concentric loops, 37
concepts, 62, 67, 100, 112, 124, 126–128, 141, 147, 187, 328
key, 189
qualitative, 128
concise overview, 148, 154
conditions, 29, 64, 83, 101, 105, 110, 114, 115, 118, 127, 131, 174, 189, 190, 213, 226, 230, 234, 235, 275, 281, 287, 288, 327, 328
anomalous, 102
complementary slackness, 255
congested, 204
ideal, 202
initial, 28, 44, 45, 49, 67, 100, 101, 113, 190, 191
input, 115
monotonicity-type, 288
operational, 127
restrictive regularity, 83
status quo, 198
configuration tree, 136, 138
configurations, 9, 114, 128, 130–140
attractor, 135
better fitness, 134, 136, 141
decentralized, 183
metastable equilibrium, 135
multiple, 131
neighbor, 135
new, 135–137
optimal, 135, 138
road, 130
root, 136
worst fitness, 136
conflict, 59, 93
warlike, 196
congestion, 23, 57, 65, 70, 77, 93, 95, 124, 136, 202–204, 306, 325
costs, 69
levels, 306, 307, 309
sharp, 70
transportation networks manifest, 56

conjectural response functions, 88
conjecture, 53, 199
connectedness, 20
connections, 9, 148, 156–159, 168, 170, 171, 177, 178, 182, 183, 198
  airport, 16
  fiber, 171
  intimate, 80
  large number of, 150, 170, 171
  missing, 4
  nodes, 179
  pairwise, 177
connectivity, 22, 173, 174, 176–178, 194
  airport, 20
  distribution, 167, 176, 178
  large, 151
  preferential, 170
  probability, 177
  properties, 148, 168, 176
  standpoint, 179
considerations, 80, 124, 156, 171
  military, 94
  technological, 79
constraints, 65–68, 71, 80, 81, 83, 88, 90, 92, 232, 235–237, 239, 243, 254, 255, 258, 274, 308, 333, 334
  definitional, 305, 306
  budget, 71, 197
  conservation, 64
  core, 88
  equilibrium-enforcing, 70
  inequality, 253
  joint capacity, 55
  linear, 201, 308
  mixed control-state, 67
  network equilibrium, 70
  nonlinear equality, 65, 72
  non-negativity, 71, 274
  physical, 175
  technological, 81, 170
  traditional conservation, 64
  transform, 279
  upper bound, 71
construction, 8, 94, 201
consumer surplus, 189–193, 198
  change, 190
  measures of, 191, 198
consumers, 69, 70, 189, 190, 271, 273–275, 287, 329
consumption, 191, 288, 329
  decisions, 275
  other-than-infrastructure, 70
contacts, 21, 23
contamination, 104, 106
  dynamic, 104
continuity, 207, 210, 239, 282, 291
continuous
  estimator, 109
  feedback control approach, 112
  time, 70, 117
contributions, 1, 59, 111, 132, 198, 227, 228, 269, 328
  fundamental, 119
control, 55, 73, 79, 80, 94, 96, 100, 101, 109, 112, 327
  algorithms, 112
  approaches, 118
  closed-loop, 112
  constraints, 83
  dynamic capacity, 225
  explicit, 67
  input, 109
  Lyapunov function, robust, 109
  paradigms, 102
  real-time, 110
  safety, 118
  system, 112
  techniques, 102
  theory, 111, 112, 117
  variables, 80, 83, 84, 88
convergence, 73, 89, 203, 207, 209, 210, 214, 216, 331
  behavior, 220
  parameters, 213
  proof of, 204, 205, 210
  test, 201
convex, 85, 87, 92, 242, 253, 260, 273, 274, 288
  hull, 208, 210
  optimization problem, 334
cooperation, 83, 158
core connectivity, 179
cost 24, 25, 27, 56, 68–70, 80, 124, 158, 173–175, 190, 194, 212, 214, 220, 287, 309, 317, 320–322
  functions, 158, 203, 205, 207, 210, 211, 214
    exponential, 308
    flow-dependent, 328
    generalized, 201, 205, 207, 214, 221

link, 334
monotonic path travel, 201
non-monotonic, 210
transaction, 273, 292, 298
turn-specific, 325
user link, 280, 295
gap, 216, 217
formulation, 211
values, 217
intermodal path, 220
minimization, 172
parameters, 206
path algorithm, 212, 218
paths, 201, 205, 209, 213
positivity assumption, 64
sensitivity, 309–311, 315, 319, 320, 325
decreasing, 325
highest, 320
lowest, 309
parameter, 308, 309
values, 311, 314, 322
structures, 223, 295, 297
cost-efficiency, 55
coupled
dynamics, 102
markets form, 60
coupling, 54–56, 72
constraints, 56, 65, 66
inter-layer, 66
Cournot model, 228
Cournot-Nash, 73
dynamic, 85
equilibrium, 62
mathematical game, 66
non-cooperative game theoretic equilibrium, 61
Cournot-Nash-Bertrand, 84, 85
agents, 88
criteria
articulate investment, 69
pre-established circuitousness, 72
critical components, 102, 187, 189, 194, 195, 198
accounts, 194
identifying, 195, 198
critical infrastructure protection, 188
cumulative distribution function, 148, 152
customers, 56, 110, 113, 224, 225, 232
charge, 223
myopic, 225
cyberinfrastructure, 57, 59
cybernetworks, 4, 59

damage, 102, 157, 171, 196, 197, 327
collateral, 93
nodes, 157
data
comprehensive, 55
distribution, 150
empirical, 154, 158
exchange, 58
extraction, 58
historical, 79
home-to-work, 175
networks, 3, 55, 326, 330, 331
real, 86, 169, 177–179, 183
record, 95
sensor, 104
decade-long investigations, 100
decays, 153, 160
definitions
basic, 8
formal, 54, 90
mathematical, 84
information theoretic, 140
of stability, 38
delays, 110, 114, 124
processing, 72
demand
amplification, 111, 114
arrivals, 224
conditions, 114
functions, 69, 189–191, 226, 229, 244, 276, 278
elastic, 69
inverse, 69
nonseparable, 69
separable, 69
step-wise, 271
time-varying, 293, 299
gap, 216, 217
learning, 223, 260
market, 271–276, 283, 291, 292, 294, 295, 298, 299, 301, 302
prices, 62, 294, 295, 301, 302
properties, 111
separability, 69
variables, 230
descent direction, 208–213

extreme point, 211
design, 7, 54, 55, 66, 69, 110, 112, 115, 118, 227
engineering, 57
infeasible, 70
internet, 94, 96
model, 70
network infrastructure, 4
verifiable, 111
destination, 8, 24, 26, 95, 123, 151, 168, 172, 175, 176, 190, 194, 204, 205, 241, 270, 280, 308, 309, 324, 328, 334
approach, competing, 168
choice user equilibrium problems, 203
flows, 308, 334
point, 241
destroy, 23, 44, 92, 196
destruction, 93, 171, 188
deterministic, 28, 29, 113, 225
laws, 28, 30
deterrence
factor, 173
function exponent, 174
detrimental dynamic effects, 111
differences, 24, 38, 49, 62, 66, 68, 90, 112, 127, 139, 149, 152, 154, 179, 189, 191, 206, 236, 243, 324
net, 189
numeric, 178
structural, 178
topologic, 178
differential
equations, 4, 30, 34, 49, 78, 81, 91, 94, 96, 100
ordinary, 7, 21
projected, 289
inclusions, 91
Dijkstra's algorithm, 195
dimensional problems, finite, 88
dimensionality, 44, 46, 136
dimensions, 31, 37, 46, 48, 49, 144, 282
socio-economic, 182
direction, 34, 78, 89, 103, 110, 154, 182, 208, 211, 214
extreme, 208, 210, 217
discontinuous travel time costs, 220
discoveries, 28, 29, 58
discrete
steps, small, 30
time
functions, 113
intervals, 225
problem, 226
disequilibrium, 65, 66, 81, 90–92
adjustment processes, 71, 78, 80
behavior, 270
capital budgeting model, 71
classical non-tatonnement, 90
dynamics, 89, 90
games, 92
impacts, 70
modeling, 89, 91
problem, 94
response, 70
states, 90
disinvestment, 94
disruptions, 93, 191, 196
maximizing, 196
disseminate network flow data, 124
distance, 16, 38, 39, 67, 68, 129, 154–156, 173, 175, 177, 194, 327
deterrence function, 173
direct effect of, 309, 334
geographical, 156, 158
mean vertex-to-vertex, 21
shortest path, 198
distribution, 13, 15, 17, 60, 62, 95, 139, 146, 150, 152, 175, 176, 178, 225, 288, 308, 320
binomial, 17
connectivity degree, 154
empirical, 95
exponential, 178
normal, 17
peaked, 15
power-law, 170, 171, 175, 176
statistical, 151
uniform, 15, 17
unique, 17
distributional assumption, 231
disturbances, 102, 110–112, 118
disutility parameters, 214, 220
dominant frequencies, 115
double-constrained specification, 174
double-dynamics theory, 270
doubly-constrained logit form, 308
dwell times, 212, 214, 220
dynamic
conflicts, 86

description, 94
game theory literature, 78
games, 2, 4, 53, 54, 77, 78, 81, 96
infinite dimensional, 78
modeling, 112
lumped, 102
models, 4, 65, 70, 88, 90, 92, 94, 96, 117, 118, 270, 275, 330, 331
continuous, 117
discrete, 112
multilayer, 66
price optimization solution providers, 224
pricing, 223–227
approaches, 226
joint, 223, 225, 228, 243, 247, 260
model, 226
strategies, 224
programming, 78
systems, 1, 7, 31, 57, 80, 91, 149, 159
competitive, 80, 92
open, 124
principles, 99, 118
user equilibrium algorithm, 208
dynamical
effects, 111
systems, 4, 28, 30, 46, 49, 51
associated projected, 291
discrete time, 31, 44
linear, 113
projected, 269–271, 289, 291
dynamics, 5, 7, 28, 30, 31, 34, 48, 49, 51, 66–68, 78, 80, 84, 85, 87, 89–91, 94–96, 101, 103–106, 110, 111, 118, 126, 127, 269, 270
internet, 96
regional, 147

economic
benefits, 69, 189
competition, 77, 86
forecasting, 93
models, 187, 188
economics, 57, 79, 126, 147, 148
evolutionary, 126
welfare, 187, 189
economy, 55, 60, 62, 65, 93, 148
competitive, 60
national, 55, 94
edge, 8, 9, 15, 16, 39, 148
effectiveness, 4, 187
effects of changing network conditions, 305, 307
efficiency, 57, 93, 168, 170, 179, 187, 194, 195
efforts, 3, 53, 94, 103, 111, 118, 119, 127, 168, 203
concerted modeling, 188
high programming, 117
mitigation, 198
national, 188
scientific, 154
eigenvalues, 40, 41
imaginary, 101
null, 101
eigenvectors, 40, 41
elastic
demand, 69, 70, 271, 276, 278
service demands, 70
user demands, 69
emergence, 4, 127, 145, 156, 158, 171
empirical values, 157
employees, 175
fulltime, 174, 175
employment, 177, 182
engineered systems, 57, 58, 99
ensembles, large, 58
entropy, 123, 139–141
environment, 56, 59, 80, 110, 118, 124, 126, 145, 146
competitive, 223, 260
data-rich, 110
decision, 73, 79
foreign economic, 80
envision, 89, 128
epidemic propagates, 23
epidemics, 21–23, 51
epidemiology, 19, 21
equality, 86, 261
equations, 28–31, 43, 45, 46, 49, 134, 138–140, 144, 145, 173, 174, 177, 191, 193, 194, 208, 210, 211, 236, 243, 261–264, 290
characteristic, 40
coupled differentials, 21
deterministic, 29
difference, 78
dynamic, 49, 81
systems of, 21, 42, 79, 262
equilibrated demand, 191
equilibrium, 24, 60, 65–68, 70, 71, 77, 80, 81, 84, 85, 89, 90, 113, 124, 190, 201, 208,

211, 214, 244–246, 255–257, 275, 276, 278, 305, 306
assignment, 212, 216
conditions, 208, 210, 269, 270, 273, 275, 276, 278, 286, 294, 295, 327
constraints, 223, 227, 228, 242, 260
demands, 246, 256
dynamic, 84
economic, 94
final, 208
flows, 278, 279, 292, 293
gap function, 205, 208
models, 53, 63, 66, 67, 183
moving, 78, 80, 84, 85, 89
path
assignment, 207, 210, 211, 220
costs, 301, 302
flows, 292, 295, 298
points, 32, 33, 46, 47, 101, 189, 206, 210
prices, 94, 189, 287
routes, 306, 307, 325
solutions, 190, 207, 210, 220, 271
state, 270, 276
time-dependent, 296
travel costs, 325
equilibrium-tending dynamics, 67, 68
equivalence, 269, 271, 278, 287
equivalent extremal problems, 87
ER models, 170
error, 12, 28, 183, 191, 234, 236, 331
absolute value, 234, 235
squared, 234–237
estimates, 13, 45, 134, 136, 167, 189, 190, 198, 201, 204, 212, 227, 228, 230, 232–239, 249, 260
parameter, 231, 232
estimation, 234, 236, 260, 325
continuous, 107
correct, 173
least-squares, 235
problem, 236, 238, 252, 258
Euler method, 291, 293–295, 299–301
evaluation, 142, 188
transit policy, 202, 203
evidence
analytical, 251
empirical, 95, 152, 154
evolution, 59, 67, 94, 100, 126, 167, 245, 246
evolutionary
universe, 126
traffic flow landscapes, 123
variational inequalities, 269, 271, 282, 287, 289–291, 297
examination, 188, 219, 319
excess, 63, 90
demand, 63, 68, 90
path costs, 63
experience, 3, 11, 73, 83, 117, 285, 326, 327, 329
personal, 326
experimentation, 103, 106, 115, 319
experiments, 103, 105, 159, 169, 173–176, 183
computational, 202
conducted, 105
exponential, 147, 150, 175
cut-off, 158
explosion, 117
functions, 155, 156, 164
expressions, 64, 70, 139, 145, 260, 264, 286, 328
extension, 57, 59, 80, 84, 89, 176, 197
externalities, 56, 190
extraneous dynamics, 106
extremal
formulation, 81
principles, 79
extremes, 57, 80

faces, 225, 226, 233
failure, 8, 22, 44, 66, 78, 93, 94, 182
cascading, 23
network, 23
system, 60
catastrophic, 108
random, 21
features
common, 147, 154
education, 59
extracted, 104
filtering, 104, 108
band-pass, 109
financial
engineering literature, 89
equilibrium problems, 288
status, 22
firms, 158, 226, 228–230, 232–234, 236, 237, 239, 243–246, 248, 251, 252, 255–257, 260–262, 264
competing, 228

set, 229
fitness, 127–131, 133, 134, 139, 140, 146
approach, 123
configurations, 127
contribution value, 132
contributions, 130, 131
criteria, 123, 128, 140, 141
function, 139
index, 134
landscape hypercube, 134
landscapes, 123, 124, 127, 131–134, 136, 140, 141, 145, 146
level, 133, 136, 140
total, 130, 132, 133, 138
values, 128–132, 135, 139, 141, 146
vectors, 134
fixed
demands, 69, 242, 252, 269, 271, 276, 278, 280, 282–284, 286, 287, 290
points, 32, 36–39, 42–48, 50, 78, 291
transfer costs, 212
flexibility, 20, 27, 110, 231
flights, 16, 18–20, 159, 232
flows, 23–26, 32, 46, 48, 56, 60, 62, 71, 90, 100, 101, 113, 130, 139–141, 172–174, 180, 276, 280, 281, 305–308, 324, 325, 329, 330
approximating, 168
arc, 26, 72
associated, 283
chain, 294–296, 301, 302
commodity, 60
economic, 177, 182
efficient, 140
energy, 54, 60
generalized transportation, 63
local, 141
origin-destination, 69, 193, 328
packet, 94
physical, 55, 326
product, 272, 291
time-varying, 66, 296
variable, 305, 306, 325
virtual, 326
forecasting, 4, 96, 159, 168, 174
formulations, 70, 71, 85, 96, 117, 173, 197, 201, 203–207, 240, 270, 289, 290, 305–308, 328
foundation, 12, 60, 66, 81, 95
disciplines, 57, 58
Fourier analysis, 103, 106
fractal, 99, 101, 128
dimensions, 104, 107, 109
models, 109
fraction, 24, 25, 27, 152
framework, 65, 124, 173, 174, 220
Frank-Wolfe algorithm, 72
free
flow, 130, 141, 145, 209, 213
networks, 16, 19, 21, 152
frequency, 115, 152, 181, 182, 306, 327
frontiers, 4, 269
fuel consumption, 309, 334
full-mixing, 22
function, 30, 56, 65, 69, 86, 102, 129, 134, 151, 152, 154, 158, 173, 189, 190, 193, 229, 234–237, 242–254, 256–258, 289, 290, 306
analytical, 117
artificial, 305, 306
boolean, 127
chain, 329
continuous, 275
decay, 154
decreasing, 230, 305, 306, 308
density, 145
deterministic, 226
deterrence, 172, 173
economic, 151, 198
engineering, 56
exponential, 154, 158
increasing, 230, 305, 309
integrable, 87
linear, 26, 230
logit, 203
monotonic, 210, 211
multinomial, 308
scaling, 109
functional
analysis results, 91
characteristics, 109
relationships, 168
functionality, 8
funds, 2, 69

game, 66, 73, 77, 78, 81, 83, 84, 87, 88, 97, 110, 111, 196
agents, 84, 87
differential, 78, 81, 85

dynamic noncooperative, 97
leader-follower, 196
many-agent mathematical, 94
multi-behavior, 83
non-cooperative, 60, 91, 197
real-world, 88
static
noncooperative, 81
Stackelberg, 74
theory, 3, 24, 57, 196, 223, 225, 227
gap function, 202, 203, 208, 210–212
gasoline, 68, 287
Gaussian distributions, 96, 231
general games, 83, 84, 88, 92
core-constrained, 88
generality, 86, 257, 264, 281, 288
generalizations, 1, 60, 84, 86, 327
generalized
costs, 69, 173, 190, 193, 194, 212, 309, 320, 322, 334
transportation
demands, 62, 72
networks, 55, 57, 60, 62, 65
generation, 2, 60, 66, 107, 138, 150
genes, 133, 145, 146
genetic analogies, 73
genotypes, 133, 145, 146
geodesic path, 9
geographical proximity, 158
geometrical inhomogenities, 102
geosciences, 147
German
commuting network, 167, 174, 176–179, 182
districts, 175, 182
infrastructure network, 182
labor markets, 177
road network, 169, 175, 180, 181, 183
Germany, 157, 169, 176, 181
globalization, increasing, 55
goods, 53–56, 60, 65, 190, 191, 194, 198, 326
governmental regulations, 55
graph, 9–11, 47, 61, 148, 149, 162, 163, 170, 194, 195, 245, 298
connected, 11
random, 11, 14
theory, 7, 9, 11, 148, 194
graphical analysis, 105
group, 12, 14, 21, 59, 102, 107, 284, 325, 329
rationality, 88
socio-economic, 325
groupings, 309, 318–320
growth, 96, 102, 159, 218
incremental, 149, 153, 170
preferential attachment-based, 169
unparallel, 187
guessing, 233, 260
guidance, 70, 187, 188

Hamilton-Jacobi-Bellman partial differential equation, 78
Hamming distance, 136
handling cost, 272, 274
hardware, 56, 94
health, 93
heuristic, 94, 201, 205, 207, 210, 214
algorithm, 210
hierarchical, 71, 83, 88, 93, 94, 151
dynamic, 80
mathematical program, 66
relationships, 330
SIM structure, 175
variational inequalities formulations, 330
hierarchy, 56, 327
horizontal, 88
high-performance computing, 59
highways, 8, 77, 126, 127, 188
Hilbert space, 82, 83, 85–92, 289
history, 93, 152, 328
horizon, finite, 226
hotel, 224, 232
hubs, 13, 15–17, 19, 20, 147, 150, 151, 154, 157, 159, 160, 177
dominant, 1, 58
emergence of, 18, 20
hierarchy of, 152, 159
structure, 171
human
behavior, 305
decision-making, 57
intervention, 56
settlement systems, 183
hurricanes, 35, 196
hypercube, 131, 132, 134
dimensional, 131, 136
hypernetworks, 54
hypotheses, 174, 175, 305–307
behavioral, 305, 306, 328
dynamic, 103

hypothetical natural disasters, 168

identification, 3, 59, 154, 269, 270, 287
implementation, 117, 142, 150, 201, 202, 207, 210, 212
  large-scale, 305–307
implications, 12, 48, 105, 110, 117, 169, 171, 325, 330
incidents, 123, 124, 136, 141
income, 65, 191
  disposable, 191
  effects, 191, 192, 198
incorporation, 168
  direct algebraic, 68
independence, relative, 141
in-depth treatment, 91
individuals, 59, 73, 196, 326
industrial dynamics, 111
industrial engineering, 99, 305
inequalities, 79, 240, 277
  quasi-variational, 302
  strict, 232
inference, 227
infinite dimensional, 270, 289
infinite time, 36, 37, 39
information, 20, 56, 86, 89, 113, 118, 124, 131, 138–140, 174, 175, 195, 196, 223, 224, 226, 227, 230, 232, 233, 239, 326
  content, 140
  customer order, 113
  discards path, 72
  exchange, 55
  flow, legitimate, 23
  infrastructures, 54, 55
  networks, 59
  paths, 111
  technology, 4, 54, 55, 60, 224, 232
  theoretic approach, 127
  traffic flow, 140
  warfare, 93
infrastructure, 3–5, 53–55, 57, 59, 66–68, 93, 94, 167, 180, 188, 269
  airline system, 331
  complete, 57
  component, 93
  decisions, 53, 54
  engineers, 3–5
  intelligent, 55
  interconnected, 188
  interdiction, 93
  internet backbone, 158
  investment, 331
  modeling, 3
  networks, 2–4, 54–56, 61, 68, 74
  planning, 331
  scholars, 59
  services, 54, 69
  systems, 3, 7, 53–55, 188, 189, 302, 327
  technologies, 54, 61, 66
  traditional, 4, 59
innovations, 20, 158
input, 109, 140, 193, 248, 333
instability, 113–115, 136
integration, 69, 70, 81, 84, 87, 118, 189
intelligent traffic system, 140
interactions, 2, 56, 57, 102, 115, 126–128, 130, 132–135, 138, 146, 148, 158, 167, 172, 198, 203, 214, 220, 327
  binary, 145
  complex, 132
  costs, 173
  parameter, 128
interdependences, 53–55, 184, 188, 202
interdependent, 54, 203
interest, 5, 28, 57, 60, 69, 79–81, 83–85, 90, 331
inter-layer dependencies, 66
inter-market, 62
intermodal, 201, 202, 205, 208, 210, 212, 220
  assignment problem, 205
  combinations, 326
  corporations, 327
  route choices, 202, 220
  test network, 214, 215
  trip equilibrium, 205
  user equilibrium, 201
Internet, 1, 80, 81, 92, 94–97, 153, 157, 159, 160, 171, 232, 326, 329, 331
  optimization models, 94
  vulnerability, 77, 81, 92, 94
interpretations, 174, 269, 271, 287
intersections, 8, 180, 305, 306, 325
intervention schemes, 58
intra-commuting, 174
inventory, 90, 110, 115, 117, 224, 229, 230
iterations, 208, 209, 211, 212, 214, 217, 218, 291
iterative scheme, 241, 291
Itô calculus, 89, 94, 96

Jacobian matrix, 69, 84
joint variational inequality, 240, 241, 243

Karush-Kuhn-Tucker conditions, 253
Kauffman's NK model, 123, 141, 145
knowledge, 7, 23, 118, 158, 159, 192, 331
Konigsberg bridge problem, 10
Kuhn-Tucker conditions, 270

label setting algorithm, 194
labour, 79, 80, 177
landscapes, 127, 128, 133, 135, 136, 141, 146
lanes, 123, 124
language, 8, 94
large scale distributed computing, 58
latency, 24–27
laws, 3, 28, 63, 172
layers, 53, 56, 65, 66, 69
Lazear's demand model, 225
leader, 87, 88, 196
learning, 8, 226, 227
left map, 323–325
length, 38, 39, 70, 144, 220
Levy-stable distributions, 96
Liapunov stability, 37, 38
lifting scheme, 108
limit cycle, 37, 48, 50, 109
limitation, fundamental, 191
limited
  energies, 109
  engagement, multiple, 93
  warfare, 77, 80, 93, 97
line, 31, 35, 44, 48, 69, 81, 87, 113, 127, 212, 217, 218
  integrals, 81, 87
  search parameters, 217
  surplus, 69, 70
linear
  assumptions, 117
  demand, 226
  predictor, best, 109
  price-demand relationships, 229
  regression, traditional, 235
linkages, 54, 64, 79
links, 9, 12, 13, 15–20, 123, 127, 147–151, 154, 158, 159, 170, 171, 180–182, 194, 195, 204, 205, 280, 281, 283–285, 293, 299, 305, 306, 309, 322–324, 333, 334
load, 22–25, 27
local
  optima, 135, 136, 141
  ordered structure, 163
locations, 35, 56, 61–63, 65, 151, 158, 310, 322, 330, 331
logit-type relations, 205
loop, 35, 36, 44, 45
loss, 95, 191, 192, 226, 327
  of generality, 257, 264, 288

machining, 101, 102, 107, 109
  dynamics, 102, 103, 109
mainframe supercomputers, 79
management, 53, 99, 123, 124, 223, 269, 305, 326
managerial independence, 56
manufacturers, 110, 271–274, 276, 283, 291, 292, 298
manufacturing, 60, 99, 100, 108, 110, 118
  applications, distributed, 112
  engineering, 7, 53, 99
  enterprises, 99, 100, 110
  processes, 99, 104, 107, 118, 119
  system 100, 101, 110–115, 117–119
map, 115, 181, 182, 310, 322
  geo-referenced, 182
  iterative, 49
  logistic, 49, 50
  right, 323, 324
market, 63, 68, 223–226, 228–230, 232–234, 236, 237, 243, 248, 260, 278
  behavior, 243
  demand, 272
  economy, 65
  electricity, 224, 228, 230
  environment influence, 224
  equilibrium, 227, 258, 264
  financial, 83
  imperfect, 190
  interdependencies, 55
  investments, 157
  multi-commodity, 68
  oligopolistic, 223
  regulations, 66
  separated, 60, 66
Markov-modulated demand model, 226
mathematical
  analyses, 1, 79
  background, 7
  descriptions, 57, 79
  formulations, 65, 70, 73, 88

foundations, 329
frameworks, 270
game theory, 79
games, 77, 78
modeling, 79, 95
models, 21, 79, 80, 201, 305, 306
perspective, 227
physics, 78
problem, 77, 228, 242
program, 70, 81, 223, 227, 242, 260
programming, 73, 80, 89, 203
representation, 59, 329
matrix, 40, 174, 229–231, 250, 251, 258
activity, 62, 65
diagonal, 240
link-path incidence, 61
positive definite, 291
maximum cost, 320
measurement problem, fundamental, 69
measures
accessibility, 168
betweenness, 179
centralization parameter, 179
convenient, 179, 210
impedance, 63
information theoretic, 123
policy, 188
qualitative, 127
mechanisms, 114, 149, 150, 159, 170
methodologies, 12, 99, 123, 168, 171, 195, 198, 244, 247, 269
methods
classical, 89
common, 169, 172
constrained descent, 71
effective, 214, 221
entropy-based, 308
finite difference, 89
inner approximation, 207
modified wavelet, 109
natural, 53
nonlinear analysis, 109
nontraditional, 72
production control, 112
wavelet denoising, 106
microsimulation, 79, 80, 83
minimization, 306, 328
modal combinations, 203
model
approximations, 45
architecture, 204
demand, 225
description, 333
fitness landscapes, 127
formulation, 72, 205, 328
games, 88
inputs, 308, 334
nonlinearities, 111
parameters, 104, 106
person movements, 214
solutions, 309, 322
transportation security, 188
travel behavior, 202
modeling, 3, 11, 54, 59, 60, 68, 77, 78, 80, 83, 92, 96, 102, 104, 109, 111, 112, 118, 134, 198, 203, 223, 269, 270
applications, 80
approaches, predictive, 118
cheap, 86
complex, 193
framework, 271
global, 107
hubs, 16
internet, 77
methodologies, 187
models, 8, 13–17, 59–61, 64–67, 71–73, 80, 81, 103, 107, 110, 111, 117, 118, 147–150, 169–171, 173–176, 195–198, 202–204, 220, 223–229, 305–308, 328–331, 333, 334
activity-based, 204
analytical, 124, 127, 129
applications of, 77, 168, 331
approximate, 193
bilevel, 92
biological, 126
chain, 284, 330
cobweb, 68
collusive, 66
combined, 203, 334
continuous, 118
data-driven, 118
demand, 228, 229, 243
descriptive, 68
final, 193
flexible intermodal, 203
flow, 330
game-theoretic, 96
gravity, 167, 170, 182

industrial dynamics, 111
informative, 55
integrated, 62
linear, 100, 109
microsimulation, 79
multi-class, 325, 328
natural conceptual, 56
nonlinear fluid, 225
predictive, 57
product manufacturing, 112
random graph, 17
regression, 107
scale-free, 171, 182, 183
simulation-based, 204
spatial interaction network, 177–179
structural, 176
supply-and-demand, 177
system-optimal, 330
tatonnement, 90
transient, 260
wavelet, 106
mode-origin-destination, 308, 333
modes, 102, 187, 202–205, 308, 309, 326, 329
combined, 202, 204
sliding, 112
of transmission, 23
modification, 96, 238
modified Tanner function, 132
modules, 204, 205
money, 56, 90, 156
monitoring, 99–103, 109, 110
monotonic cost functions, 202, 205, 210, 220
monotonicity, 203, 207, 210
motivation, 126, 223
movement, 134, 183, 191, 196
multi-layer, 56, 65, 69, 71, 72
model, 66
multi-objective routing algorithm, 204
multiple harmonics, finite, 104
multi-scale representations, 109

Nash,
equilibrium, 24–27, 226, 234, 244, 245, 252, 255, 257, 260, 262
demands, 228, 234–236, 238, 243, 244, 246, 249, 252, 255–258, 261, 262, 264
prices, 247, 250, 253
problem, 239, 252
revenues, 248, 251, 254
solution, 26, 27, 197
Nash-Cournot game theory, 329
national infrastructure systems, 55
national productivity, 59
national science foundation, 2, 4, 59, 119
national strategy, 188
nature, 30, 54, 88, 132, 134, 152, 158, 180, 183, 196
neighborhood, 36, 38, 134, 138, 270
network, 1–3, 7–12, 14–17, 19, 20, 22–27, 50, 51, 53–55, 58, 59, 126–132, 134–136, 138–141, 147–154, 156–159, 167–171, 174–183, 194–198, 214–216, 306, 307, 325–327, 329
agents, 58
analysis, 194
applications, 13
approaches, 149, 168, 174
arcs, 198
assignment, 220
attributes, 194
BA model, 177
capacity expansion, 73
changes, 198
characteristics, 20
conditions, 305, 307
configurations, 58, 130
connectivity, 154, 176
controller, 58
damage, 196
demands, 193, 197
design, 60, 66, 69
disruptions, 196
distribution, 150
equilibrium, 203, 325, 331
evolution, 19
failure, 171
flows, 65, 127, 140
game agents, 93
games, 23
growth, 170
improvement, 331
interdiction, 80
Internet infrastructures, 157
large, 4, 27, 135, 306, 307
layers, 53, 72
links, 61, 331
match, 182
models, 2, 148, 149, 162, 182, 188, 328, 329
multilayer, 56, 270

multimodal, 201, 202, 205
neural, 73, 104, 112, 149
optimization models, 187
parameters, 149, 183
performance, 170
perspective, 152, 182
physical, 2, 175, 180, 182, 189, 305, 306
planar, 171
power distribution, 22
problems, 329
properties, 169, 175, 179, 183
random, 1, 15, 20, 147, 148, 152
real, 11, 17, 19, 21, 179
representation, 324
scale-free, 20, 22, 153, 171, 175, 183
science, 1–4, 7–9, 11, 12, 22, 53, 57, 58, 92, 147, 302, 325
security, 188
segments, 130
simulation, 169
states, 131
structures, 21, 58, 176, 180, 183, 270, 272
topologies, 23, 177, 194
NK model, 123, 127, 145, 146
nodes, 8–17, 19, 20, 22–24, 42, 44, 58, 62, 138, 148–151, 154, 157–159, 162, 163, 170, 171, 174, 179–182, 194, 195, 205, 283, 305–308, 333
abundance of, 147, 150
breakdown, 20, 22
configuring, 17
connected, 171, 174, 194
degenerate, 42, 44
destination, 283, 287
fourth-tiered, 283
number of, 16, 147, 150
sequence of, 308, 333
single, 16, 171
well-connected, 182
noise, 106, 108, 119
nonchaotic behavior, 50
nonconvex, 72, 88
nonconvexity, 65, 72
noncooperation, 24, 83, 92
non-cooperative case, 66
noncooperative conflicts, 86
non-cooperative dynamic equilibrium, 85
nondecreasing function, 24
nonlinear, 30, 42, 99, 102–104, 106, 112, 115, 117, 118, 260
complementarity, 305, 307
problem, 82, 203, 308
dynamics, 7, 28, 50, 100, 107, 118, 119
applications of, 100, 103
monitoring of, 99, 101, 103, 105, 107, 109, 111, 113, 115, 117, 119, 121
principles, 105, 109, 118
effects, 115
functions, 30
models, 100, 103, 119
optimization, 308
phenomena, 103
processes, 101, 103, 108, 111
programming, 78, 302
response, 115
science, 2–5, 7, 53, 57, 58, 302
signals, 107, 108
stochastic dynamics, 110
systems, 30, 105
nonlinearities, 93, 96, 102, 104, 109
non-tatonnement dynamics, 90, 91
non-technical description, 307
non-traditional hierarchical formulations, 88
notations, 40, 61, 62, 68, 229, 271, 272, 292
numerical, 86, 210, 271, 291, 296
dynamic, 269, 290
example, 73, 290, 292, 299
methods, 83, 96
observations, 210
results, 73, 86

objective function, 81, 196, 234, 235, 237, 239, 242, 258, 260, 274
objects, discrete, 144
observations, final, 152
operating
conditions, 101, 114, 115
costs, 292, 298
operations, 7, 56, 69, 92, 101, 110, 113, 118, 141, 188, 195
operators, 62, 86, 327
optima, global, 73
optimal control problems, 4, 71, 80, 83, 88, 90–92
optimality, 253, 327
conditions, 82, 252, 273, 274, 276, 278, 305, 328
optimization, 2, 77, 93, 96, 242, 329
models, 94

problems, 83, 235, 241, 242, 273, 327
orbit, 35, 39, 101
order parameter, 139
origin-destination couples, 175
origin-destination generalized cost, 308
origins, 55, 151, 172, 190, 194, 205, 241, 270, 280, 308–310, 323, 328, 333, 334
outcomes, 68, 83, 101
outputs, 79, 193, 221
overlapping perspectives, 57

packets, 92, 94, 95
pairs, 8, 61, 177, 241, 280, 281, 283, 285, 287, 288, 292–295, 299, 301, 302, 306, 312, 320
origin-destination, 72, 241
zone, 309, 322, 324
paradox, 26, 70, 93, 123, 124
parameters, 30, 45, 46, 50, 73, 104, 109, 115, 141, 173, 179, 183, 226–228, 230, 231, 233–236, 238, 242, 244, 245, 248–250, 256–260, 262
partial fitness tree, 137
path 9–11, 20, 26–29, 31, 44, 61–63, 72, 149, 151, 179, 201, 214, 216, 218, 280, 281, 284, 285, 287, 292–295, 299, 301, 302
assignments, 206, 208, 218
calculations, 220
costs, 212, 290
enumeration, complete, 72
flows, 269, 281, 282, 284, 287, 288, 290, 291, 295
generation, 212, 217, 218
identities, 72
interactions, 214, 220
iterations, 218
intermodal, 202, 204, 212
physical, 180
shortest, 9, 20, 149, 179, 194, 195, 211
spatiotemporal, 205
subnetwork-specific, 62
unambiguous, 81
patterns, 104, 106, 114, 129, 136, 176, 322
peaks, 128, 132, 146, 182
performance, 27, 59, 99, 102, 106, 110–114, 118, 218, 331
perishable products, 226, 228, 230
persistence, 157, 159, 160
person, 8, 9, 12, 21, 24, 28, 212, 214, 308, 333
movements, 212, 220
perspective, 4, 20, 54, 55, 80, 126, 181, 199, 326
perturbation, 28, 29, 42, 113
phase, 94, 210–213, 216–218
diagram, 31, 35, 44
point, 31, 32
transitions, 11, 57, 58
physics, 59, 183
planning, 4, 54, 55, 67, 92
economic, 92–94
players, 86, 88, 110, 196
Poisson, 150, 151
distribution, 150, 162
models, 95, 96
policies, 53, 68, 77, 94, 111, 187, 188, 226
population, 16, 21, 23, 49, 67, 126, 127, 145, 152, 153, 182
possible configurations, 128, 130, 135, 138, 146
total number of, 130, 140
post-bifurcation dynamics, 103
power, 27, 94, 115
computational, 169
distribution characteristic, 19
explanatory, 68, 96
generation, 56
law, 147, 150–156, 158, 164, 169
practitioners, 224, 325, 330, 331
prediction, 28, 79, 80, 118
preferential
attachment, 19, 150, 153, 157, 172, 174, 176
connectivity behaviour, 182
price, 62, 68, 73, 90, 189, 191, 223–226, 228–233, 236–238, 244–251, 256, 260, 275, 276, 279, 287, 292, 327, 329
of anarchy, 26
elasticities, 69, 229
functions, 234, 237, 239, 244
historical, 236, 237
notation, 293
observations, 248, 249
optimal, 223, 228, 239
parameters, 233, 237
sensitivities, 229, 230
variations, 226
price-demand, 235
parameters, 228, 232, 234, 235, 237, 241, 242, 258, 260
relationship, 223, 227, 229–231, 239, 260

pricing, 224, 225, 330, 331
models, 224
optimal congestion, 190
option, 83, 89
problem, 224, 228, 241
dynamic, 227, 228, 238, 239
strategy, 225
prioritization scheme, 198
priority, 199
routes, 218, 219
transit signal, 214
probability, 11, 14–17, 129, 131, 134, 138, 140, 150, 152, 154, 226
distribution, 139, 140, 148, 227
functions, 196
problems, 9, 11, 23, 65, 69, 73, 77, 79–83, 87–89, 197, 198, 205, 210, 223, 225–228, 234–237, 241–243, 260, 270, 271, 326, 327, 329–331
best-response, 237, 241, 242, 252, 255, 260
extremal, 81, 92
process, 17, 79, 99, 100, 102–104, 106, 110, 117, 118, 124, 127, 134, 137, 141, 189, 218, 235, 329
control, 109, 118
dynamics, 103, 104, 106
monitoring, 103, 105, 106
producers, 69, 70, 189, 190
production
cost, 272, 273
quotas, 330
rates, 113
systems, 110, 112, 118
products, 101, 102, 110, 113, 116, 117, 228, 271–276, 297, 329
customized, 110
homogeneous, 271
inner, 82, 85, 91, 282
physical, 329
price-insensitive, 287
profiles, 109, 126, 146
complex, 146
surface, 109
traffic pattern, 126
profits, 225, 228
expected, 224, 227
negative, 63
programming languages, common, 73
projected dynamical systems theory, 270, 271
projection, 78
methods, 89
minimum norm, 68, 70, 90
operator, 289
projective dynamics, 67, 68, 90, 91
projects, 68, 188, 191
promotions, 88, 225
propagation, 107, 203
properties
structural, 169, 174, 181
topological, 7, 8, 11, 13, 15, 20, 22, 23, 149
protocols, 94, 96
proxy, 173, 175

quadratic rule, 40
quantifiers, 109
quantities, 226, 273
queuing, 202

railroad, 326
random
breakdowns, 21
graphs of Erdös, 12, 13
inspections, 196
utility theory, 205, 308
vaccinations, 22
variable, 139, 140
randomness, 15, 124, 201
rank 135, 152, 175
order, 136
size rule, 152, 153
rate, 49, 68, 90, 95, 107, 214, 224–226
rational learning behavior, 226
recognition, 1, 93
recurrence relations, 30
recurrent self-organizing behaviors, 58
redesign, 110, 118
redistribution, 159
regenerative effects, 102, 103
region, 49, 57, 68, 83, 93, 123, 128, 140, 158, 168
regulation, 327
relaxation scheme, 241
release time, 112
repeat phase, 213
repellers, 32, 35, 37
replenishment, 224, 227, 228
replication, exact, 79
representation, 8, 65, 104
graphical, 162, 164

wavelet, 107
research, 12, 50, 80, 83, 97, 103, 110, 118, 119, 141, 167, 169, 182, 202, 203, 221, 225, 227, 302, 328, 329, 332
resemblance, 172
resonance conditions, 102
response, 101, 104, 105, 112, 113, 115–117, 153
continuous-time, 100
linearized, 116
muted, 226
retailers, 224, 271–276, 278, 283, 291, 292, 298, 299, 329
revenues, 226, 229, 237, 241, 244–247, 249, 273, 274
total, 229, 233
revolutionary changes, 126
rewiring, 15, 16, 170
ridership, 202, 203, 219, 220
risk, 23, 56, 188, 327
road, 8, 23, 24, 123, 126–128, 144, 145, 181, 182
infrastructure, 183
network, 123, 127, 130, 134–136, 141, 168, 169, 181, 182, 197, 305, 306, 308, 310, 320, 322, 324, 325, 333
toll, 331
traffic, 305, 306, 330
transportation, 142, 307
robustness, 20, 112, 157, 170
roots, 46, 47, 73, 136
route, 8, 23, 24, 26, 27, 50, 61, 123, 128, 182, 183, 202–204, 206, 218, 219, 305, 306, 308–325, 328, 330, 331, 333, 334
choices, 202–204, 217
combinations, 325
costs, 206, 320, 334
flows, 331, 334
maps, 331
maximum number of, 310–312, 315, 322, 324
number, 219
pattern, 309
segments, 324
structure, 324
router, 157
routing, 23, 24, 26, 61, 62, 330
decisions, 329
frequency, 179
model, 330
optimal regulated, 26
paths, 182
principle, 62
routing-stop-percentage, 213, 218
rudimentary examples, 127
rush orders, 110, 112

saddle point, 33, 34, 37
scale, 7, 16, 19, 21, 22, 109, 150, 152, 154, 312, 320
double logarithmic, 153
free networks, 148
freeness, 2
geographical, 168
invariance, 101
log-log, 175, 312
multiple, 99, 101
natural, 58
perturbations, 135
scale-free, 148
scaling factor, 173
scarce funding, 187
scholars, 1, 2, 59, 306
scholarship, 4, 59
science, 1, 57, 152
regional, 2, 167, 168, 172
screening models, 79, 80
seasonable demand, 297
segments, 126–136, 138–141, 144, 179
blocked, 130, 131, 145
congested, 136
interacting, 128
neighboring, 128, 129
number of, 138, 140
selfish routing, 26, 27
self-similarity, 58, 94, 95, 101
seller, 224–228, 232, 241, 242
selling horizon, 224, 227–229, 232, 233, 241, 243, 248
separability assumptions, 83
separable case, 69
separation, 106, 147, 149, 168
sequence, 78, 81, 105, 291, 324
service coverage, 93
services, 55, 65, 94, 128, 191, 327, 331
levels of, 127, 128, 130, 138, 140, 141, 145
settings, 58, 84, 224, 225, 230, 239, 297
2-firm-2period, 260
capacitated, 260

competitive, 225
dynamic, 60, 70, 84
multi-product, 225
real world, 69
shipments, 326–329, 331
characteristics, 327
freight, 326
grain, 327
moving container, 196
natural gas, 326
shippers, 327–329
behavior of, 327, 328
shop floors, 99, 117, 118
short term market changes, 111
shortest path algorithm, 180
signal, 68, 103, 104, 106, 107, 109, 113
features, 106, 107
linear, 108
measured, 104–107
nonstationary, 105
representation, 104, 106
separation, 104, 106
timing, 203
vibration, 106
signal-state relationship, 107
simple deterministic differential equations, 96
simplicial decomposition, 208
simplicity, 128, 230, 235, 297
SIMs, 168, 169, 171–179, 182, 183
double-constrained, 175, 183
unconstrained, 169, 173
simulated annealing, 73, 74
simulation, 4, 20, 29, 86, 93, 94, 117, 124, 162, 163, 201, 217, 218, 248, 250, 251
agent-based, 73
experiments, 155, 156
models, 29, 169
results, 214
single-scale networks, 153
sinusoidal oscillations, 115
small world, 12, 15, 20, 147, 148, 152
effect, 7, 11–14, 19, 20
networks, 153
social network, 14
social
cost, 24–26
networks, 1–4, 8, 9, 12–14, 22, 59, 149
relationships, 8
sciences, 148, 167, 172
socioeconomic characteristics, 56
soft-thresholding, performing, 106
software, 56, 72, 73, 80, 95, 112
antivirus, 23
tools, 60, 66
solutions, 10, 11, 26, 27, 30, 71, 72, 82, 83, 85, 86, 100, 206–208, 210, 216, 240–242, 258, 269, 270, 276, 277, 279, 288, 289, 293–295, 300, 301, 306–310, 319–325
algorithm, 205, 207
congested, 323, 324
efficient, 190
explicit, 297
global, 73
local, 88
new, 208, 209, 211, 212
numerical, 89, 94
optimal, 24, 26, 27, 136
simultaneous, 236
system-optimal, 328
unique, 207
spectrum, 57, 83, 270
continuous broadband, 101
spirals, 35, 37, 42, 44, 48
splits, 22, 47, 157
squares, 21, 87, 247
stability, 37, 38, 42, 45, 47, 57, 79, 81, 89, 91, 113
assymptotic, 37, 38
Stackelberg-Cournot-Nash mathematical game, 66
Stackelberg-Cournot-Nash-Bertrand, 84, 87, 91
dynamic games, 84
Stackelberg-like agent, 92, 94
state, 28, 30, 31, 35, 45, 67, 68, 78, 80, 83, 84, 86, 88, 90, 117, 129–131, 140, 202, 226, 239, 273, 276, 327
agencies, 188
conditions, 29, 230
dynamics, 71, 83, 84
flows, 66
initial, 45
observable, 90
space, 45, 89, 101, 117
stationary, 94
transitions, 117
unknown, 107
unsteady, 114
variables, 67, 90, 104

statement, 63, 70, 325
static
  equilibria, 60, 84
  models, 112, 275, 306, 331
  network design, 72
stations, 68, 112, 113, 116
statistical
  characteristics, 147
  measurement, 150
  mechanics, 59, 73, 183
  trends, 93
  viewpoint, 154
statistics, 57, 147, 178, 179, 227
step, 209, 228, 233–237, 239, 242, 243, 260
  length, 208–210, 213, 214, 217
  response, linearized, 117
step-wise demand, 298
stochastic, 81, 94, 96, 205, 225, 226
  assignment problem, 204
  differential equations, 96
  term, intrinsic, 96
stochasticity, 96, 124, 126
strategies, 187, 196, 210, 212
  non-cooperative decision-making, 59
  optimal pricing, 225, 237, 238
structural
  changes, 4, 79
  indicators, 178
  limitation, 171
structure, 7, 8, 11, 22, 82, 88, 106, 167, 168, 173, 174, 176, 178–182, 227, 230, 297, 306, 307, 324, 325, 334
  centralized, 183
  chain, 111, 278
  free, 7
  functions, 109
  institutional, 188
  mathematical, 81, 83, 89, 107
  physical, 182
  scale-free, 171
  step-wise, 296
  topological, 7, 22, 179
subject, 71, 79, 171, 273, 274, 305, 306, 327, 330
subnetworks, 1, 58, 62
subroutes, 324
subset, 101, 138, 140, 208
  bounded, 288
  nonempty, 82
substitution, 67, 87
subsystems, 59, 79
subtasks, 103, 104
subtracting, 10, 69
suburbs, inner, 324
subvector, 229
sufficiency, establishing, 253
sum, weighted, 309, 329
supernetwork, 269, 271, 278, 283, 293–296, 299, 301, 326
  constructed, 276
  equivalence, 282, 288
  representation, 291, 293, 299
  topology, 297, 298
supplier requirements, 113
supplies, 272
  contaminate water, 56
  separated, 55
surface
  finish, 102
  integrity, 102
  patches, 109
  undulations, 102
surveillance equipment, 140
survivability, 20, 128
symmetry, 84, 233
  requirements, 83
  restrictions, 70, 81, 85
synergism, 54, 55
synthesis, 328
systems, 4, 7, 28, 30–32, 34, 35, 37, 39–42, 44–46, 48–51, 53–57, 59, 60, 90, 99–101, 103, 104, 109–112, 114–118, 126, 129, 138–141, 157
  architecture, 57
  behavior, 115, 126
  biological, 125, 145
  chaotic, 28, 49
  characterization, 119
  competitive, 60, 78, 80
  evolutionary, 126
  large, 59
  linear, 30, 39, 42, 101, 105
  natural, 28
  oscillating, 37
  parameterized, 48
  physical, 45, 100, 113, 156
  random, 129
  real-world, 99, 100
  solar, 28

of systems, 53–57, 59, 60
system-wide measures, 219

tatonnement dynamics, 90
Taylor, 42, 43
expansion, 44
TCP, 95
techniques, 57, 72–74, 104–106, 138, 188, 193, 194, 198, 330
approximate, 193, 194, 198
complex network science, 187
measurement, 59
numerical analysis, 94
state space reconstruction, 101
traditional pricing, 224
technological factors, 160
technologies, 4, 123, 127, 140, 142, 148, 269
advanced information, 59
communication, 86
data-logging, 110
technology investments, 198
telecommunication networks, 326, 327
telecommunications, 55, 57, 70, 148
telephony, classical, 95, 96
terminology, 79, 84
terrorism, 187, 188
tests, 105, 174, 175, 202, 214, 217, 220, 243
real-world, 214, 220
statistical, 154, 160
threats, 171, 188
external, 187, 198
hostile, 198
threshold, 106, 177
throttling, 23
tiers, 60, 271, 276, 283
time
horizon, 113, 291, 293, 295, 299–301
long, 100, 102, 159
periods, 95, 135, 136, 191, 202, 223, 226–229, 232, 234, 236–239, 241, 243–249, 257, 260, 262–264, 305, 307
time-dependent, 201, 205, 269, 270, 288
demands, 271
equilibria, 295
modeling framework, 271
network conditions, 204
transportation equilibrium problems, 288
timeframes, distinct, 270
time-varying, 78, 275, 288
demands, 269, 287, 291
least-cost paths, 202
realities, 202
traffic conditions, 202, 203
tolls, 327, 328
tools, 99, 110, 117, 118, 126, 160, 174, 224, 225
top tier prices, 278
topological
characteristics, 11, 20
connotations, 128
topology, 1, 14, 15, 20, 22, 23, 34, 45, 50, 51, 128, 148, 158, 159, 176, 295
hub-based, 21
of networks, 8, 151
visualizations, 178
tornado, 45
total
flow, 24, 25, 305, 306, 309, 334
latency, 26, 27
track, 29, 112
trade, 65, 187
facilitation, 198
organizations, continental, 327
trade-off, 80, 226
traffic, 8, 23, 27, 126–129, 134, 135, 140, 141, 193, 194
alert, 56
assignment, 201, 270
conditions, 55, 124, 127, 134, 135, 190, 206
congestion, 123
equilibria, 82, 306, 331
flows, 124–129, 134–136, 138, 140–142, 193, 198, 305, 307
incidents, 141
jams, 123, 129
landscapes, 136, 141
networks, 23, 124, 127, 129, 131, 140, 141, 325, 331
patterns, 126
properties, 134
simulation, 202, 220
systems, 129
tragic events, 187
trajectories, 29–31, 33–39, 42, 44–46, 48–50, 78, 89, 90, 100, 101
transactions, 65, 274–276
transfers, 202, 203, 205, 216, 329
transient dynamic pricing problem, 233
transient effects, 94

TRANSIMS model, 124, 204
transit, 308, 309, 334
transition, 11, 31, 163
translations, 292–295, 300–302
transmission, 23, 95, 329
network, 22, 149
transportation, 3, 57, 82, 167, 174, 187, 188, 193, 221, 271, 326, 328
agencies, 193
analogy, 241
benefits, 190
costs, 191, 193
demand model, 193
economics, 187
facilities, 192, 193, 196, 198
field, 239, 241
freight, 307
generalized, 53, 54
group, 124
infrastructure, 187, 188, 201
models, 191, 193
networks, 7, 8, 60, 123, 128, 168–171, 174, 188, 189, 191, 193–196, 198, 287, 288, 327, 328, 330
planning agency, regional, 190
policies, 201
problems, 241, 326
projects, 190
science, 168, 269, 270
security, 187, 198
services, 65, 191
systems, 2, 7, 56, 187, 188, 198
technologies, 201
terminology, 327
travel, 8, 16, 23, 24, 27, 126, 180, 241, 270, 309, 325
behavior, 202, 270
costs, 201, 205, 280, 285, 294, 301, 302, 309, 320, 325, 328
demands, 280, 281, 285
disutilities, 280, 285, 292, 296
forecasting, 325
time, 23, 24, 123, 128, 173, 175, 204, 220, 241, 305, 306, 334
traveler, 9, 201, 205, 206, 215, 217, 241, 270, 285, 309, 325, 328, 331
behavior types, 216
benefit, 331
groups, 214
preferences, 220
travelling, 136, 308, 334
tree, 136
binary, 138
trips, 50, 126, 202, 204, 205, 208, 214–217
automobile, 90, 204
home-to-home, 168
intrazonal, 309
non-collaborative, 126
two-period horizon, 226

uncertainty, 56, 139–141
unconstrained case, 257, 258, 264
uncut chip thickness, 102
unification, 81, 269, 271
unified constraint set, 288
unilateral decision, 206
uniqueness, 78, 79, 81, 82, 91, 207, 270
United States, 1, 7, 16, 20, 53, 77, 79, 93, 94, 99, 108, 183, 188
unpredictability, 28
unused capacity, 229, 244, 245, 255, 257, 261
user-equilibrium, 322, 328, 334
user-optimization, 270, 287
users, 27, 53, 54, 58, 124, 127, 189–194, 204, 270, 280, 327
utility, 3, 56, 174, 191, 307, 327, 328

vaccination technique, 23
validate, 214, 220
values
absolute, 152, 291
average, 144
computed, 145
economic, 188, 189, 193, 198
final, 217
initial, 49
numerical, 128
observed, 237
random, 133
unambiguous, 69
variables
decision, 78, 226, 230
dummy, 87
independent random, 231
variational inequalities, 65, 70, 78, 80, 83, 86–89, 94, 201, 203, 207, 223, 239, 240, 269, 270, 273–278, 280, 282, 286, 289, 305, 327
differential, 77, 79, 80, 91

finite-dimensional, 83, 269, 270, 273, 282, 290
formulation, 65, 202, 269, 276, 278, 286
infinite dimensional, 82, 83, 89
joint, 239, 241, 242
lower level, 260
nonlinear, 88
problem, 80, 82, 270, 282, 291
single joint, 240
variations, 84, 102, 103, 109, 114, 115, 141, 146, 197, 212
coupled nonlinear, 103
equilibriating, 192, 198
vector, 30, 34, 62, 130, 229, 240, 272, 281, 287, 288
continuous, 291
field, 32, 34, 35, 37, 291
nonnegative, 63
of path costs, 290
process parameter, 101
vector-valued disutility function, 85, 91
vertex, 8, 9, 15, 131–134, 136, 140, 149
bad fitness, 140
connectivity, 154
generic, 149
vibrations, 102, 103
virus, 22, 23
visualization, 31, 162, 163, 179, 181
vulnerability, 19, 21, 50, 80, 94, 157, 196

Walrasian price equilibrium problems, 287
warfare, 92, 93
wars, 77
Watts-Strogatz model, 16
wavelet transforms, 104
wealth, 223, 232
websites, 232, 241
weight, 159, 194, 206, 327
weighted combination, 134
welfare, 190
well-defined subproblems, 78
winner, 58
work-in-process, 113, 116
workplace, 23, 123
workstations, 113, 114, 116, 117
world networks, 21

x-axis, 34, 154

zones, 124, 174, 214, 305–309, 333

Early Titles in the
INTERNATIONAL SERIES IN
OPERATIONS RESEARCH & MANAGEMENT SCIENCE
Frederick S. Hillier, Series Editor, *Stanford University*

Saigal/ *A MODERN APPROACH TO LINEAR PROGRAMMING*
Nagurney/ *PROJECTED DYNAMICAL SYSTEMS & VARIATIONAL INEQUALITIES WITH APPLICATIONS*
Padberg & Rijal/ *LOCATION, SCHEDULING, DESIGN AND INTEGER PROGRAMMING*
Vanderbei/ *LINEAR PROGRAMMING*
Jaiswal/ *MILITARY OPERATIONS RESEARCH*
Gal & Greenberg/ *ADVANCES IN SENSITIVITY ANALYSIS & PARAMETRIC PROGRAMMING*
Prabhu/ *FOUNDATIONS OF QUEUEING THEORY*
Fang, Rajasekera & Tsao/ *ENTROPY OPTIMIZATION & MATHEMATICAL PROGRAMMING*
Yu/ *OR IN THE AIRLINE INDUSTRY*
Ho & Tang/ *PRODUCT VARIETY MANAGEMENT*
El-Taha & Stidham/ *SAMPLE-PATH ANALYSIS OF QUEUEING SYSTEMS*
Miettinen/ *NONLINEAR MULTIOBJECTIVE OPTIMIZATION*
Chao & Huntington/ *DESIGNING COMPETITIVE ELECTRICITY MARKETS*
Weglarz/ *PROJECT SCHEDULING: RECENT TRENDS & RESULTS*
Sahin & Polatoglu/ *QUALITY, WARRANTY AND PREVENTIVE MAINTENANCE*
Tavares/ *ADVANCES MODELS FOR PROJECT MANAGEMENT*
Tayur, Ganeshan & Magazine/ *QUANTITATIVE MODELS FOR SUPPLY CHAIN MANAGEMENT*
Weyant, J./ *ENERGY AND ENVIRONMENTAL POLICY MODELING*
Shanthikumar, J.G. & Sumita, U./ *APPLIED PROBABILITY AND STOCHASTIC PROCESSES*
Liu, B. & Esogbue, A.O./ *DECISION CRITERIA AND OPTIMAL INVENTORY PROCESSES*
Gal, T., Stewart, T.J., Hanne, T./ *MULTICRITERIA DECISION MAKING: Advances in MCDM Models, Algorithms, Theory, and Applications*
Fox, B.L./ *STRATEGIES FOR QUASI-MONTE CARLO*
Hall, R.W./ *HANDBOOK OF TRANSPORTATION SCIENCE*
Grassman, W.K./ *COMPUTATIONAL PROBABILITY*
Pomerol, J-C. & Barba-Romero, S./ *MULTICRITERION DECISION IN MANAGEMENT*
Axsäter, S./ *INVENTORY CONTROL*
Wolkowicz, H., Saigal, R., & Vandenberghe, L./ *HANDBOOK OF SEMI-DEFINITE PROGRAMMING: Theory, Algorithms, and Applications*
Hobbs, B.F. & Meier, P./ *ENERGY DECISIONS AND THE ENVIRONMENT: A Guide to the Use of Multicriteria Methods*
Dar-El, E./ *HUMAN LEARNING: From Learning Curves to Learning Organizations*
Armstrong, J.S./ *PRINCIPLES OF FORECASTING: A Handbook for Researchers and Practitioners*
Balsamo, S., Personé, V., & Onvural, R./ *ANALYSIS OF QUEUEING NETWORKS WITH BLOCKING*
Bouyssou, D. et al./ *EVALUATION AND DECISION MODELS: A Critical Perspective*
Hanne, T./ *INTELLIGENT STRATEGIES FOR META MULTIPLE CRITERIA DECISION MAKING*
Saaty, T. & Vargas, L./ *MODELS, METHODS, CONCEPTS and APPLICATIONS OF THE ANALYTIC HIERARCHY PROCESS*
Chatterjee, K. & Samuelson, W./ *GAME THEORY AND BUSINESS APPLICATIONS*
Hobbs, B. et al./ *THE NEXT GENERATION OF ELECTRIC POWER UNIT COMMITMENT MODELS*
Vanderbei, R.J./ *LINEAR PROGRAMMING: Foundations and Extensions, 2nd Ed.*
Kimms, A./ *MATHEMATICAL PROGRAMMING AND FINANCIAL OBJECTIVES FOR SCHEDULING PROJECTS*
Baptiste, P., Le Pape, C. & Nuijten, W./ *CONSTRAINT-BASED SCHEDULING*
Feinberg, E. & Shwartz, A./ *HANDBOOK OF MARKOV DECISION PROCESSES*: *Methods and Applications*
Ramík, J. & Vlach, M./ *GENERALIZED CONCAVITY IN FUZZY OPTIMIZATION AND DECISION ANALYSIS*

Early Titles in the
INTERNATIONAL SERIES IN
OPERATIONS RESEARCH & MANAGEMENT SCIENCE
(Continued)

Song, J. & Yao, D./ *SUPPLY CHAIN STRUCTURES: Coordination, Information and Optimization*
Kozan, E. & Ohuchi, A./ *OPERATIONS RESEARCH/ MANAGEMENT SCIENCE AT WORK*
Bouyssou et al./ *AIDING DECISIONS WITH MULTIPLE CRITERIA: Essays in Honor of Bernard Roy*
Cox, Louis Anthony, Jr./ *RISK ANALYSIS: Foundations, Models and Methods*
Dror, M., L'Ecuyer, P. & Szidarovszky, F./ *MODELING UNCERTAINTY: An Examination of Stochastic Theory, Methods, and Applications*
Dokuchaev, N./ *DYNAMIC PORTFOLIO STRATEGIES: Quantitative Methods and Empirical Rules for Incomplete Information*
Sarker, R., Mohammadian, M. & Yao, X./ *EVOLUTIONARY OPTIMIZATION*
Demeulemeester, R. & Herroelen, W./ *PROJECT SCHEDULING: A Research Handbook*
Gazis, D.C./ *TRAFFIC THEORY*
Zhu/ *QUANTITATIVE MODELS FOR PERFORMANCE EVALUATION AND BENCHMARKING*
Ehrgott & Gandibleux/ *MULTIPLE CRITERIA OPTIMIZATION: State of the Art Annotated Bibliographical Surveys*
Bienstock/ *Potential Function Methods for Approx. Solving Linear Programming Problems*
Matsatsinis & Siskos/ *INTELLIGENT SUPPORT SYSTEMS FOR MARKETING DECISIONS*
Alpern & Gal/ *THE THEORY OF SEARCH GAMES AND RENDEZVOUS*
Hall/*HANDBOOK OF TRANSPORTATION SCIENCE - 2nd Ed.*
Glover & Kochenberger/ *HANDBOOK OF METAHEURISTICS*
Graves & Ringuest/ *MODELS AND METHODS FOR PROJECT SELECTION: Concepts from Management Science, Finance and Information Technology*
Hassin & Haviv/ *TO QUEUE OR NOT TO QUEUE: Equilibrium Behavior in Queueing Systems*
Gershwin et al./ *ANALYSIS & MODELING OF MANUFACTURING SYSTEMS*
Maros/ *COMPUTATIONAL TECHNIQUES OF THE SIMPLEX METHOD*
Harrison, Lee & Neale/ *THE PRACTICE OF SUPPLY CHAIN MANAGEMENT: Where Theory and Application Converge*
Shanthikumar, Yao & Zijm/ *STOCHASTIC MODELING AND OPTIMIZATION OF MANUFACTURING SYSTEMS AND SUPPLY CHAINS*
Nabrzyski, Schopf & Węglarz/ *GRID RESOURCE MANAGEMENT: State of the Art and Future Trends*
Thissen & Herder/ *CRITICAL INFRASTRUCTURES: State of the Art in Research and Application*
Carlsson, Fedrizzi, & Fullér/ *FUZZY LOGIC IN MANAGEMENT*
Soyer, Mazzuchi & Singpurwalla/ *MATHEMATICAL RELIABILITY: An Expository Perspective*
Chakravarty & Eliashberg/ *MANAGING BUSINESS INTERFACES: Marketing, Engineering, and Manufacturing Perspectives*
Talluri & van Ryzin/ *THE THEORY AND PRACTICE OF REVENUE MANAGEMENT*
Kavadias & Loch/*PROJECT SELECTION UNDER UNCERTAINTY: Dynamically Allocating Resources to Maximize Value*
Brandeau, Sainfort & Pierskalla/ *OPERATIONS RESEARCH AND HEALTH CARE: A Handbook of Methods and Applications*
Cooper, Seiford & Zhu/ *HANDBOOK OF DATA ENVELOPMENT ANALYSIS: Models and Methods*
Luenberger/ *LINEAR AND NONLINEAR PROGRAMMING, 2nd Ed.*
Sherbrooke/ *OPTIMAL INVENTORY MODELING OF SYSTEMS: Multi-Echelon Techniques, Second Edition*
Chu, Leung, Hui & Cheung/ *4th PARTY CYBER LOGISTICS FOR AIR CARGO*
Simchi-Levi, Wu & Shen/ *HANDBOOK OF QUANTITATIVE SUPPLY CHAIN ANALYSIS: Modeling in the E-Business Era*

**** A list of the more recent publications in the series is at the front of the book ****