Statistics for Social and Behavioral Sciences

Statistical Confidentiality

Principles and Practice

Springer

Statistics for Social and Behavioral Sciences

For other titles published in this series, go to
http://www.springer.com/series/3463

George T. Duncan · Mark Elliot ·
Juan-José Salazar-González

Statistical Confidentiality

Principles and Practice

Springer

George T. Duncan
Carnegie Mellon University
Santa Fe, NM 87505, USA
gtduncan@gmail.com

Mark Elliot
University of Manchester
Manchester, UK
mark.elliot@manchester.ac.uk

Juan-José Salazar-González
University of La Laguna
La Laguna, 38271 Tenerife, Spain
jjsalaza@ull.es

ISBN 978-1-4419-7801-1 e-ISBN 978-1-4419-7802-8
DOI 10.1007/978-1-4419-7802-8
Springer New York Dordrecht Heidelberg London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Get together with statisticians and you may see a T-shirt emblazoned, “In God we trust, all others bring data.” And, beyond doubt, statisticians are bent on getting data. Indeed they are fully employed in the full gamut of sample surveys, government censuses, observational studies, and clinical trials. Observe another T-shirt bannered, “Top 10 Reasons to be a Statistician,” listing, “Estimating parameters is easier than dealing with real life.” Surely this self-effacing humor makes the contrary point: that, while statisticians do deal with probability models and their parameters, what really fascinates them is how it all relates to life, and especially the real uncertainties of life. They are intrigued with how data can yield information to reduce these uncertainties, and so enable better decisions to be made. Our concern in this book is that data relevant to issues in the public interest, such as a person’s medical record or criminal history, are often highly sensitive. Obtaining and using such data forces us to realize that there is tension between, on the one hand, the desire of the individual for a full and free private life and, on the other hand, the needs of the broader community for information that might, say, improve health care or reduce crime. Statistical confidentiality is pivotal in resolving this tension.

This book on statistical confidentiality is written for all involved with personal and proprietary data from empirical studies. Various roles require an understanding of statistical confidentiality. Here are some instances drawn from issues concerning assessment of a drug rehabilitation program:

- You are a researcher. Your undertaking is to seek rich and convincing evidence to assess benefits of the program, both to addicts and to the community.
- You are a statistician. You design a survey of addicts and link the results to administrative data from the program.
- You are a data steward. Your responsibility is to take the statistician’s data and build a database useful to researchers *and* acceptable to privacy advocates.
- You are a privacy advocate. You express qualms about the researcher and statistician matching drug-use records to the addict’s personal finances and any criminal behavior.
- You are a respondent to the survey. In principle you are happy to provide data about yourself for the good of society, but you are also concerned about what happens to that data once you have handed it over: will your privacy be respected; will your confidentiality be maintained?

So what exactly is statistical confidentiality? Simply put, it is the stewardship of data to be used for statistical purposes. Stewardship, as expressed in statistical confidentiality, is an active embrace of responsibility for both protecting data and ensuring its beneficial use. Explicitly it requires proper practices for both providing and restricting access to data products.

Getting and using data just to support public policy analysis costs a lot of money. Lane (2003) makes this point: "Billions of taxpayer dollars are spent in supporting the collection and dissemination of federal, state and local data, billions of dollars are spent in data analysis, and this, in turn, both informs scientific understanding of core social science issues and guides decision in how to allocate billions of dollars in social programs." Privacy and confidentiality concerns do not come in dollars, pounds or Euros, but quantifying these concerns via Google search on May 19, 2010, yielded more than 1.42 billion hits on "privacy", more than 19.7 million hits on "confidentiality", and more than 55,000 hits on "statistical confidentiality" (including the quotation marks in the query). Just looking at the first few of these hits on "statistical confidentiality" leads us to information on a United Nations work session in Geneva, Switzerland, to a discussion of pertaining laws in France from the National Institute for Statistics and Economic Studies, to the US Federal Register discussion of the Statistical Confidentiality Order, to a research site of the Computational Aspects of Statistical Confidentiality Project based in the Netherlands, to testimony before the US Congress regarding confidentiality and coordination among statistical agencies, and to a discussion of confidentiality in the Japanese 2000 Census of Population.

The issues cited above illustrate the scope of statistical confidentiality. By absorbing the ideas in this book, you will gain understanding in both the principles and practice of this important field that can benefit your work:

- As a researcher, you will understand why an agency holding statistical data does not respond well to approaching them saying, "Just give me the data; I'm only going to do good things with it," and appreciate why you need to learn about what motivates statistical confidentiality and how it works in practice.
- As a statistician, you will incorporate the requirements of statistical confidentiality into your methodologies for data collection and analysis.
- As a data steward, you are caught between those eager for data and those who worry about confidentiality in its dissemination. Fortunately, using the tools of statistical confidentiality you will progress toward satisfying both groups.
- As a privacy advocate, you will comprehend how confidentiality can be protected even though statistical data are, and should be, made available.
- As a respondent, you will have a better understanding of why your data are needed, how they will be used, and how they will be protected.

We have organized this book into eight chapters. In Chapter 1 we motivate and define the study of statistical confidentiality, laying out the dilemma of data stewardship organizations (we will call them DSOs; important examples are statistical agencies) in resolving the tension between protecting data from snoopers and providing data to legitimate users. We identify the stakeholders in the statistical process, show why statistical data are so useful and in such demand, and explain why DSOs are so concerned about confidentiality. We explore the concept of disclosure risk in terms of an attack by a data snooper and show the basic ways statistical confidentiality can be protected. In Chapter 2 we lay out the fundamental concepts of statistical confidentiality, develop conceptual models of disclosure risk and data utility, identify ways risk can be assessed and controlled, and explore the types of attack that a data snooper might mount. From a rational decision-making perspective, Chapter 3 presents the methodologies of disclosure risk assessment, including a variety of useful metrics for risk assessment. Chapter 4 gives techniques for statistical disclosure limitation of aggregate data, specifically data in tabular form. We present an appropriate definition of disclosure-limited tabular output. We develop deterministic methods, especially through mathematical programming, and stochastic methods, such as cyclic perturbation, for statistical disclosure limitation of tabular data. Chapter 5 gives techniques for disclosure limitation of microdata, that is, data in original record form. We affirm the value of microdata and clarify what users need from such data. We identify the concerns that a DSO has in satisfying the ethical, pragmatic, and legal considerations that motivate their confidentiality promises to data providers. We point out the characteristics of microdata that make it vulnerable to confidentiality attack, and explore various masking methods. We introduce the idea of synthetic data, that is, data that are stochastically generated from a model inferred from the source data. In Chapter 6 we give measures of the impact of disclosure limitation on data utility, and develop the methodology of R-U confidentiality maps and their empirical analog. Chapter 7 provides restricted access methods as a body of administrative procedures for disclosure control. We explore the issues a DSO must face in deciding who can have access, where can access be obtained, what analysis is permitted, and what modes of access should be allowed. Finally, Chapter 8 explores the future of statistical confidentiality. We address a number of important questions: Will privacy and statistical confidentiality have new meanings? Who will care about statistical data? What new forms of DSO will develop? Will statistical data remain valuable? Will there be new issues for statistical confidentiality? Will there be new forms of data snooping? What new strategies of disclosure limitation should be developed?

Plainly, statistical confidentiality is a large and important issue of concern. Research and work in statistical confidentiality is growing rapidly, as is the number of people working on this problem. Many people are screaming that their privacy and confidentiality are vanishing. It is the job of DSOs to provide as much quality data as possible without violating confidentiality laws and promises. This book provides a reader with a comprehensive understanding of the principles and practice of statistical confidentiality. It balances methods and ideas with specific examples. We have written it to be accessible to those just entering the field. Much of the material requires no specific background in mathematics or statistics. Those sections which are more technical are prefaced with explanations of the general ideas without complicated equations.

Our thanks go to our many colleagues who helped us comprehend the need for new ways of dealing with statistical confidentiality, collaborated in our research on this topic over many years, and inspired this book.

Santa Fe, New Mexico, USA	George T. Duncan
Manchester, UK	Mark Elliot
La Laguna, Spain	Juan-José Salazar-González

Contents

Chapter 1
Why Statistical Confidentiality?

> *As a society we are judged by how we treat those at the dawn of life, those in the sunset of life and those in the shadows of life.*

Confirming statistical confidentiality as vital to the stewardship of personal data, the United Nations set out Principle 6 of its Fundamental Principles of Official Statistics: *Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.*[1] Data stewardship is active on two fronts:

- protecting entrusted data by providing confidentiality
- assuring its beneficial use by researchers and policy analysts

The US Census Bureau assures, *Data Stewardship is the formal process we use to care for your information—from the beginning, when you answer a survey, to the end, when we release statistical data products. Data Stewardship goes beyond the law to ensure that any decisions we make will fulfill our ethical obligations to respect your privacy and protect the confidentiality of your information ... The Data Stewardship program ensures that we protect the information you provide, while enabling us to release high quality information about our population and the economy.*[2]

In this chapter, we address the following questions:

1. What is statistical confidentiality? (Section 1.1)
2. Who are the stakeholders in the statistical process and what are the stakes for them? (Section 1.2)
3. What is the dilemma in statistical confidentiality? (Section 1.3)
4. What is the utility of statistical data and why is it in such demand? (Section 1.4)
5. Why are Data Stewardship Organizations (DSOs) so concerned about confidentiality? (Section 1.5)

[1] unstats.un.org/unsd/goodprac/bpabout.asp

[2] www.census.gov/privacy/data_stewardship/partnership_and_trust.html

G.T. Duncan et al., *Statistical Confidentiality*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-7802-8_1,

6. What is it about high-quality statistical data that raises confidentiality concerns? (Section 1.6)
7. What is disclosure risk and who is the data snooper who might want to attack statistical confidentiality? (Section 1.7)
8. How can statistical confidentiality be protected? (Section 1.8)

Building on this base, Chapter 2 lays out the concepts of statistical disclosure limitation. Chapter 3 presents the methodologies of disclosure risk assessment. Chapter 4 gives techniques for statistical disclosure limitation (SDL) of aggregate data, specifically data in tabular form, while Chapter 5 gives techniques for SDL of microdata, that is, data in original record form. Chapter 6 gives measures of the impact of SDL on data utility. Chapter 7 describes restricted access methods as a body of administrative procedures for disclosure control. Finally, Chapter 8 explores the future of statistical confidentiality.

1.1 What Is Statistical Confidentiality?

First, what do we mean by *confidentiality*—for now leaving aside the adjective "statistical"? Broadly, confidentiality reflects a complex amalgam of societal values about information privacy, secrecy of personal information, and autonomy of the individual. But more specifically, confidentiality is a status accorded to information about a person. Under confidentiality, the party holding the information is bound by an implicit or explicit promise that it be protected from unauthorized or inappropriate access and usage. Fienberg (2005) defines confidentiality this way: "Broadly, a quality or condition accorded to information as an obligation not to transmit that information to an unauthorized party."

The term "confidentiality" is used in many different contexts, ranging over religious confessionals, communications from patient to doctor, employment recommendations, and asserted prerogatives of officials to garner input shielded from outside monitoring. Regardless of context, confidentiality is a promise that is made to the provider of the information by the receiver and current holder of the information. There are two overarching principles to confidentiality: information should be (1) reserved exclusively for intended purposes, and (2) used only by authorized individuals.

Promise keeping is not the only aspect of the ethical basis for confidentiality. Other ethical considerations include a respect for the individual's autonomy, a desire not to cause individuals embarrassment, and a view that society functions better when confidentiality is considered to be a human right. Additional considerations, especially pragmatic and legal ones, motivate concern for confidentiality. We will explore this in detail in Section 1.5.

We view statistical confidentiality as a body of principles, concepts, and procedures that permit confidentiality to be afforded to data, while still permitting its use for statistical purposes. This is not an easy task, and so arises the need for this book.

1.2 Stakeholders in the Statistical Process

To understand how statistical confidentiality functions, we must appreciate that the various interests among researchers, statisticians, data users, and data providers rarely align and so must be reconciled by the responsible organization. Data stewardship organizations manage the process of data capture, storage, integration, and dissemination. DSOs include statistical agencies, national statistical offices, data archives, trade associations, unions, credit bureaus, and health information associations. Many DSOs support research and policy analysis based on statistical methods.

To illustrate the DSO's problem let us consider medical data. Here is a listing of just some of the stakeholders in such data:

- Patients
- Physicians
- Family members
- Hospital administration
- Public health officials
- Insurance companies
- Governments at all levels
- Employers
- Health care researchers
- Educators
- Journalists
- Marketers
- Law enforcement
- Litigants

Patients choosing a hospital might want statistical information on hospital mortality rates. Insurance companies seek statistical information on costs per patient. Health care researchers want objective, high-quality statistical data about things such as hospital admissions of teenage drug users and how they relate to socioeconomic conditions of the patients. Marketers want statistical data on trends regarding the willingness of patients to accept generic alternatives to prescription drugs. Law enforcement wants statistical data on admissions to emergency facilities for gunshot wounds. Hospital administrators need medical information for billing purposes and quality assurance purposes. Journalists want to provide analyses to their readers. Family members want information about their loved ones. Given the variety of these expressed needs, a DSO faces a real and growing tension between protecting privacy and satisfying the demand for information.

1.3 The Data Stewardship Organization's Dilemma

Astonishing advances in technology for computing and telecommunications have dramatically altered how we make decisions. In our domain of statistical confidentiality, these advances have also precipitated a dilemma: On the one hand, technology has boosted the perceived benefits of statistical information. Consider that three-quarters of Americans polled[3] said it was very important or somewhat important for doctors and hospitals to use electronic records instead of paper. On the other hand, people have heightened anxiety about whether confidentiality is, or even can be, provided. The same poll revealed that nearly the same number of respondents said they were not confident that their computerized records would remain safe from prying eyes. What is the meaning of these conflicting developments for the DSO, which must make decisions about confidentiality and data access?

On the one hand, the benefits of statistical information about people became apparent as early as 1830 when Adolphe Quetelet examined factors influencing social phenomena such as crime, marriage, and suicide (see Stigler, 1986). Illustrating the importance of government data in public health, John Snow used data in 1854 from the Registrar of Deaths to show the link between cholera and contaminated water in London. Today the social statistics enterprise is huge. For example, the Organization for Economic Co-Operation and Development (OECD) gathers voluminous statistical data from at least 30 countries on topics ranging from agriculture to health to globalization to transport, and disseminates data products to an estimated 6 million users in over 100 countries.

On the other hand, increasing availability of massive amounts of data about individuals has also fueled confidentiality concerns. Here are two examples:

- In 2006 US Congress members and the Federal Communications Commission raised concerns about the sale on several Internet sites of customers' wireless and landline phone records, including the date, time, and length of calls placed by consumers.
- Controversy ensued when in 1998 the Icelandic parliament granted the biopharmaceutical company deCODE genetics the right to construct an electronic database of the country's health records together with genetic information on some 65% of the Icelandic population and a genealogy that for most stretches back 1000 years.

Because a DSO is a broker between the providers of data and the users of data, it is caught in the colliding paths of information demands and confidentiality concerns. DSOs broadly encompass all organizations that capture, store, integrate, and disseminate information—the CSID (Capture Storage Integration Dissemination) data process (Duncan, 2004). In the statistical realm we are concerned with data capture that involves surveys (for example the American Community Survey), censuses (such as the UK Census 2011), and administrative procedures (such as providing the required information to obtain a medical license in Spain). Today, data storage is ubiquitously electronic and measured in terabytes. Electronic storage facilitates the integration of records across disparate databases, for example a health economics

[3] http://www.npr.org/templates/story/story.php?storyId=103362165

study might integrate an individual's medical record with their employment history. Data dissemination makes a data product such as a collection of tables or a microdata record file available to a data user.

The largest and most prominent of DSOs are those whose primary function is statistical. These include government statistical agencies (e.g., the US Bureau of the Census), national statistical offices (e.g., Statistics Canada), and cross-national statistical offices (most prominently, Eurostat—the Statistical Office of the European Community).

Another grouping of DSOs is the data libraries and data archives (e.g., the National Data Archive for Child Abuse and Neglect at Cornell University[4]), which typically are not heavily involved in primary data collection, but do assemble statistical databases and make them available.

A third grouping of DSOs includes those with rather specialized functions. These include trade associations (e.g., National Association of Manufacturers), unions (e.g., in Spain, Unión General de Trabajadores, UGT), credit bureaus (e.g., Experian), marketing data firms (e.g., Nielsen//NetRatings, Marketing Opinion and Research International (MORI), which is a large market research firm in Great Britain), and health insurance information agencies (e.g., the Health Insurance Industry Benchmarking Association).

A fourth grouping of DSOs is organizations for which the collection of data for statistical purposes is not part of their primary business, but which are increasingly been called on to enable the use of such data that they do collect for administrative purposes to be used for statistical research. For example, government departments in the United Kingdom provide data for the Neighbourhood Statistics Service (NeSS).

Many DSOs are experts in what is needed to acquire high-quality data from individuals, households, and establishments. A DSO serves the operational and decision needs of the community at large in the case of government statistical agencies, specialized researchers and policy analysts in the case of data libraries and data archives, and various client constituencies in the case of the specialized DSOs. Regardless of their mission, the challenge to each DSO is to manage the critical and sensitive data under its care so that they can be used to their fullest capacity.

Also, whatever the DSO and its particular function, the DSO's professional staff has the responsibility of developing and implementing confidentiality and data access policies. The DSO has to be concerned with what data should be afforded confidential status and how it can be protected—specifically how the risk of disclosure, which is access by unauthorized parties, can be limited (see, for example, Duncan, 2001).

The Secretary General of the European Commission affirmed the value of statistical confidentiality to the European Parliament: "Official statistics must be

[4] www.ndacan.cornell.edu/

produced and disseminated according to common standards guaranteeing compliance with the principles of impartiality, reliability, objectivity, scientific independence, cost-effectiveness and statistical confidentiality."[5] According to this statement, the DSO must fulfill conflicting standards: one standard that requires an organization to *provide* quality data products at low cost, and a second mandating that the organization maintain statistical confidentiality. This second standard is essentially an imperative to *not provide* information that can be tied to individuals. Having one mandate to provide data and a second to not provide data emphasizes the vexing nature of the DSO's predicament. How can one have a duty to reveal and yet a duty to hide? Nonetheless, the DSO has an opportunity to develop creative ways to resolve this inherent tension and so provide the factual basis for decision making—a critical component that benefits the common good—while simultaneously demonstrating that the individual's concerns about confidentiality can be respected.

To summarize, given that providing complete access to statistical data and maintaining confidentiality are ineluctably opposing goals, key staff within DSOs face a predicament whose resolution requires a grasp of the principles and practice of statistical confidentiality. In Section 1.4 we discuss why statistical data are of such value and in Section 1.5 we discuss why DSOs are so concerned about statistical confidentiality.

1.4 The Value of Statistical Data

Statistical data are the foundation for empirical analysis, providing the factual information needed to guide policy and decision making and to enable a better scientific understanding of how our world works. Empirical analysis addresses questions about human conditions and needs. For example, empirical analysis using statistical data can address such questions as the following:

1. Can middle-class neighborhoods absorb low-income families under a housing mobility policy without leading to neighborhood decline through out-migration of affluent families?
2. How can chest pain be diagnosed as myocardial ischemia in women?
3. What combination of pharmacological and psychosocial interventions works best for bipolar mood disorder?
4. Will lowering tax rates increase tax revenue?
5. Will restricting immigration of the technically skilled raise employment opportunities for poor citizens?

To illustrate the need for statistical data, let us consider issues akin to Question 1 above. In 1999, the Urban Institute sponsored the Symposium on Section 8 Mobility

[5] eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52005DC0217:EN:NOT

and Neighborhood Health (Section 8 is a tenant-based housing assistance program in the United States). The report on this symposium[6] identified the need for the following statistical data:

1. Data to map Section 8 locations, identify clusters of Section 8 recipients, and assess market conditions in neighborhoods where these clusters are located.
2. Local data to understand and address potential behavior problems and neighborhood consequences.
3. Data for determining the availability of below-Fair Market Rent housing units in neighborhoods of different types regionwide, and relating clusters of Section 8 housing to the distribution of affordable housing.

Those who work to bring such statistical data to bear on research questions include policy analysts, sociologists, epidemiologists, psychiatric researchers, and economists. They are employed variously by universities, think tanks, trade associations, unions, marketing and consulting firms, corporations, and government departments and agencies.

Answering such questions requires statistical data of high quality, which necessitates that it be accurate, detailed, and comprehensive. Often it needs to be at the level of the individual person. Further, it often must identify what is happening in both time and space. So the data need to be geographically specific and longitudinal. These criteria are explored in the next section in conjunction with their implications for statistical confidentiality. Here we note that the demand for statistical data has led to the development of a variety of accessible data sets that are widely useful. Here are two examples that illustrate what is readily available on the web and suggest how useful they can be:

1. *US Census Bureau PUMS files—Public Use Microdata Sample*[7] Recognizing that "Because of the rapid advances in computer technology and the increased accessibility of census data to the user community, the Census Bureau has had to adopt more stringent measures to protect the confidentiality of public use microdata through disclosure-limitation techniques," this sample provides individual and household-level data from the 2000 Census long form. These data provide comprehensive information on social and economic characteristics of the people as well as physical and financial aspects of housing.
2. *The WHO Statistical Information System*[8] Here the World Health Organization presents the most recent statistics since 1997 of 50 health indicators for WHO's 192 Member States. A highlight: "While some countries are making

[6] www.urban.org/url.cfm?ID=309465

[7] www.census.gov/population/www/cen2000/pums.html

[8] www.who.int/whosis/en/

progress and achieve greater equality in child survival chances within the country, the general picture is that little progress has been made during the last decade."

1.5 Why Are DSOs Concerned About Statistical Confidentiality?

Concerns about confidentiality have presumably been an issue ever since one person learned something about another. Indeed, the 4th century BC Hippocratic Oath mandated for physicians, "All that may come to my knowledge in the exercise of my profession or in daily commerce with men, which ought not to be spread abroad, I will keep secret and will never reveal." In the Hebrew Oath of Asaph, the practitioner is admonished: "Ye shall not disclose secrets confided unto you."[9] The Bible acknowledges there is "a time to keep silence and a time to speak" (Ecclesiastes 3:7). Notwithstanding this long history of concern, many of today's information privacy worries are new, and are driven by both technological advances that have lowered costs and societal changes, especially those involving mobility and population growth.

A large, mobile population has raised the demand for personal information. Health care is no longer the sole province of a local physician providing continuing care to a patient over dozens of years. Instead, health care must be provided wherever the patient may happen to live, work, and play—whether in Madrid or Adelaide—and so benefits from databases of electronic records. Commerce is not just the purview of the local merchant, but extends globally and works through instant credit checks against electronic databases and electronic tracking of shipping.

1.5.1 A Difficult Context for a DSO

Over the years, statisticians and other professionals in DSOs have contributed to methodologies and practices that enable the collection of quality statistical data. Yet today the analysis that is required to answer many questions of societal importance is threatened because the faucet to useful statistical data may become so restricted that nothing but a trickle emerges. The statistical community is in this crisis of data access for four inter-related reasons: privacy worries, confidentiality concerns, a changing legal and social context, and an increased sensitivity to social impact. We now examine each of these reasons, which all essentially relate to whether a data provider can trust a DSO.

[9] http://www1.umn.edu/phrm/oaths/oath5.html

1.5.1.1 Privacy Worries

Privacy is closely aligned with confidentiality, but distinct from it. Privacy is linked to the desire of individuals to control how they are presented to others. Also, as those others see it, privacy is not being intrusive into the lives of people. Information privacy has been defined to encompass "an individual's freedom from excessive intrusion in the quest for information and an individual's ability to choose the extent and circumstances under which his or her beliefs, behaviors, opinions, and attitudes will be shared or withheld from others"(Duncan et al., 1993).

Not surprisingly, the appropriate scope of privacy is hotly debated. Much of this debate is outside the scope of this book. For example, we will not discuss how much privacy employees should have in using e-mail in the workplace, or how much authority government should have to search your laptop's files. Yet, much of the debate about privacy *is* within the scope of this book. For example, we will address privacy issues in areas such as research use of personal medical information, the sharing of statistical data in the United States between the Social Security Administration and the National Institute on Aging, and providing public-use data files from the Korean National Fertility Survey.

A quick search of the web reveals the breadth of public concern about privacy in almost every area where information is collected:

1. The Electronic Frontier Foundation shows how secret codes inserted in the output of color printers can be used to trace their source.[10]
2. BBC News raises concerns about Radio Frequency Identification (RFID) tags on consumer goods.[11]
3. CBS News reports on personal data on airline passengers being given to the government for testing a computerized background-check project.[12]

1.5.1.2 Confidentiality Concerns

As with privacy worries, confidentiality concerns are ubiquitous, appearing wherever information is shared. Think of these problems:

1. Duke University Medical Center reports that HIV patients in rural areas say they're afraid to seek treatment because they fear breaches of confidentiality by their medical providers.
2. In the United Kingdom the Equal Opportunities Commission seeks pay records to check for sex bias, but does this require employee permission?

[10] http://hardware.slashdot.org/article.pl?sid=08/02/15/1612226

[11] http://news.bbc.co.uk/2/hi/technology/6691139.stm

[12] http://ssc.sagepub.com/cgi/content/abstract/23/4/401

3. Perhaps most bizarrely, in the United Kingdom, confidentiality researchers, themselves, have been denied access on confidentiality grounds to the data needed to test whether their confidentiality protection systems work!

1.5.1.3 Changing Legal and Social Context

To see the relevance for privacy and confidentiality of a changing legal and social context we need only consider how the events of September 11, 2001, altered government stances on the capture, storage and integration of personal data. For example, in 2006, the US government obtained access to telephone records from major telecommunications companies and engaged in widespread data mining. One company, Qwest, refused on privacy grounds to provide records.[13]

More broadly than the consequents to the tragic events of 9/11, policies and practices related to privacy and confidentiality have been profoundly affected by an array of societal changes:

1. The formation and expansion of the European Union has forced consideration of incompatibilities in legislation regarding data flows.
2. The explosive growth of e-commerce has strained protection of financial and other information on individuals and led to concerns about identity theft.
3. The perhaps even more explosive growth of web-based social networking (see, e.g., Facebook[14]) has raised privacy issues particularly in how much personal information a young person may share with those who are essentially strangers.
4. The ubiquity of blogs and online message boards has raised concerns about anonymous defamatory postings on the one hand and protection of freedom of speech on the other hand.

Not surprisingly given their public policy importance, pro-privacy groups maintain websites that address and often advocate positions on such issues. For example:

1. Electronic Frontier Foundation www.eff.org/
2. Electronic Privacy Information Center http://www.epic.org/
3. Privacy International www.privacyinternational.org/
4. Privacy Rights Clearinghouse www.privacyrights.org

1.5.1.4 Sensitivity to Social Impact—"Group Harm"

A criticism has been levied at some social research projects that the statistical outcomes might present certain groups unfavorably. Sensitivity to this issue has led to recommendations for the nature and scope of research projects that might be

[13] www.nytimes.com/2006/05/12/washington/12cnd-phone.html?ex=1305086400&en=16b1c1d512d1d04b&ei=5088&partner=rssnyt&emc=rss

[14] www.facebook.com

conducted. For example, in the United States a 1999 National Bioethics Advisory Commission report was reviewed by a multi-agency Working Group that noted:

> The Working Group is persuaded that further work is needed to identify appropriate methods for consultation with representatives of groups. While it is appropriate for IRBs [Institutional Review Boards] to consider risk of group harm even for studies conducted with anonymised samples, the current regulations do not provide a mechanism for DHHS [U.S. Department of Health and Human Services] to require IRB review of such studies. The Working Group agrees with Recommendation 18, which states that risk of group harm should be disclosed during the process of obtaining informed consent for research.[15]

For large-scale statistical surveys, similar concerns about group harm have led to suggestions that data not be released at levels of detail sufficient to provide information about particular groups. This raises serious questions about who should make such decisions, since statistical information that may be harmful to a group in one sense may also benefit the same group in another. For example, if a study found that a group was differentially subject to alcoholism, that finding might be taken as stigmatizing to the group and hence harmful, yet the study might also lead to additional resources being directed toward the group for prevention and treatment.[16] In such cases, who is to decide that the cost is greater than the benefit or vice versa?

Furthermore, releasing statistical data for small geographical areas (e.g., block-level data) has been criticized because of the potential to use these data to disadvantage individuals as members of certain localized groups. How should this concern be addressed? More broadly the issue is not just releasing data for small geographical areas, but releasing any group-differentiated data (for example data differentiated by demographic variables, such as age, sex, and race). The problem is that the very usefulness of the statistical data (that it differentiates and shows relationships between constructs) is also its pitfall (that it highlights and underlines prejudices)—use and abuse are two sides of the same coin.

1.5.2 Providing Data and Protecting Confidentiality

In fact, both access to high-quality data and confidentiality protection are essential, and no DSO can long function without accomplishing both. Yet data access and confidentiality are not immediately compatible for at least three reasons:

1. Provision of high-quality data is a responsibility to the users of data—the researchers and analysts. The DSO must construct useful statistical products for

[15] http://aspe.hhs.gov/sp/hbm/execsum.htm

[16] To illustrate, studies done by the Indian Health Service in the United States have shown that the rate of alcoholism-related deaths for American Indian youth between the ages of 15 and 34 is over 12 times that of the general population. This has led to the Indian Health Service funding some 400 alcoholism and substance abuse programs, which provide treatment and prevention services to both rural and urban communities.

a broad spectrum of data users. Confidentiality, on the other hand, is a responsibility the DSO holds to the providers of the data, the individuals, households, organizations, and firms. The DSO must assure data providers that participation is helpful and not harmful. Thus the DSO is serving two masters—each with conflicting interests and concerns.

2. We can "solve" the confidentiality problem by not making available any data product at all. But, that completely undermines the primary function of a DSO. Conversely, we compound the confidentiality problem by releasing all the data.
3. We can't solve the confidentiality problem by simply deidentifying the data records—removing obvious identifiers such as name or social security number. Just for fun, take a look again at the quotation that began this chapter. It has been deidentified—the author's name has been removed. How hard is it to reidentify this record? Try searching the web and see if you can do it. (Hint: the quote may not be exact—just as data may not be.) An answer is at the end of the chapter.

If effective, the DSO resolves the inherent tension between satisfying the desires of data subjects for confidentiality and the desires of data users for high-quality data. In resolving this tension, they broker the relationship between the data providers and the data users.

1.5.3 Consequences of a Confidentiality Breach

DSOs are often reluctant to provide certain useful data products to their clients because they cannot ensure confidentiality in the face of an inference attack by a data snooper (Garfinkel et al., 2002). With a successful attack, the confidentiality of the data would be compromised. What are the consequences of such a compromise? Empirical assessment of consequences is impossible because there is essentially no public history of such compromises. In the case of the United Kingdom, for example, Elliot and Dale (1999) observe that, "There has been no known attempt at identification with the 1991 SARs [Samples of Anonymised Records]—nor in any other countries that disseminate samples of microdata." Certainly an operative word in this statement is "known," so this should not be taken to mean that DSOs should relax their vigilance about confidentiality. But it does mean that DSOs have not—perhaps because of their vigilance—had to deal with the public consequences of a confidentiality breach.

This lack of experience for DSOs is quite unlike bad events for other organizations. Airlines, for example, have experienced crashes with loss of life, and so know something of the impact of disasters on operations and reputation. In spite of a lack of direct experience, DSOs, and especially National Statistical Offices, are convinced that the publicity of a confidentiality breach would seriously damage their reputation as an honest broker and so their ability to collect statistical data. As the Eurostat Manual of Business Statistics (2005) puts it, "Statistical confidentiality is necessary in order to gain and keep the trust of those required to respond to statistical surveys."

1.5.4 What Motivates a DSO to Provide Confidentiality?

As noted in National Research Council (2005), there are three fundamental motivations for DSOs to develop policies and methods to protect confidentiality: (1) it may be required by law or regulation, (2) it may compromise the ability of a DSO to do its job (which is a pragmatic consideration), and (3) a confidentiality breach may violate ethical norms.

1.5.4.1 Legal Requirements and Fair Information Practices

There is a body of legislation that deals quite generally with personal privacy information issues. In many European countries such issues are referred to under the label of *data protection.* Much of this legislation is based on the statement of Fair Information Practices put forth in 1980 by the Organization for Economic Cooperation and Development. These principles have evolved and been simplified through the work of various groups, particularly governmental agencies in the United States, Canada, and Europe. The resulting fair information practice codes have five core principles of privacy protection, notification, consent, respondent access, data integrity, and enforcement:

1. *Notification*

   Respondents should be given notice of a DSO's information practices before any personal information is collected from them. Generally notification will include identification of:

   a. the DSO collecting the data
   b. the uses to which the data will be put
   c. potential recipients of the data
   d. the nature of the data collected and the means by which it is collected
   e. whether the provision of the requested data is voluntary or required, and the consequences of a refusal to provide the requested information
   f. the steps taken by the data collector to ensure the confidentiality, integrity, and quality of the data

2. *Consent*

   To the extent practicable, potential respondents should be able to choose whether to provide personal information. Generally there are two different types of consent mechanisms, *opt-in* and *opt-out*:

   a. Opt-in mechanisms require positive steps by the respondent to allow the collection and use of information.
   b. Opt-out regimes require positive steps by the respondent to prevent the collection and/or use of such information.

3. *Respondent access*

   A respondent has access when he or she can view their own data, and can contest accuracy and completeness.

4. *Data integrity*
   Data should be accurate and secure. Security requires both managerial and technical measures to protect against loss and unauthorized access. Managerial measures include administrative procedures that limit access to data and ensure that those individuals with access do not use the data for unauthorized purposes. Technical security measures to prevent unauthorized access include encryption in the transmission of data, use of password protection, and the storage of data on secure computers.
5. *Enforcement*
   Fair information practice acquires teeth only through an adequate enforcement mechanism. Enforcement can be based upon three different mechanisms:

   a. Self-regulation. This generally involves making compliance with a code of fair information practices a condition of membership in an industry or professional association and external audits to verify compliance. A self-regulatory mechanism should provide a means to investigate and redress complaints from individual respondents.
   b. Private remedies. Legislation could authorize respondents harmed by a DSO's information practice to file suit for compensatory or punitive damages.
   c. Government sanctions. Legislation could provide for civil or criminal penalties for violation of fair information practice.

Many DSOs are under specific or general legal constraints to afford confidentiality to the data they steward. Because of the importance of this enforcement mechanism, we will provide a more detailed discussion of its various implementations. Demonstrating the worldwide scope of such legal requirements, McCaa and Odinga (2001) observe that, "of the 54 member-states of the International Monetary Fund's General Data Dissemination System, almost all are bound by law to respect the privacy of individuals and maintain statistical confidentiality of the information collected." Spain, for example, promulgated the "Ley Orgánica 15/1999 de Protección de Datos de Caracter Personal"[17] to regulate statistical confidentiality issues. Also note Regulation 831/2002 of the Commission of the European Communities which regulates data access for scientific purposes.[18]

The prominent piece of US legislation dealing with statistical confidentiality was enacted for the Census Bureau under Title 13 (see Cecil, 1993). Prior to 2002, most other US federal agencies did not have specific legislation protecting statistical confidentiality. Since the passage of the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), all federal agencies are required to protect their data obtained under a pledge of confidentiality for exclusively statistical purposes from being disclosed in identifiable form. CIPSEA defines *identifiable form* as any representation of the data that permits the identity of the respondent to be reasonably inferred by either direct or indirect means. Some agency staff have expressed the belief that this will result in a higher level of confidentiality and that this may encourage respondents to participate in surveys and that agencies

[17] "Boletín Oficial del Estado," number 298, day 14 December 1999, pp. 43088–43099

[18] heur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:133:0007:0009:EN:PDF

can thereby avoid disputes about withholding data from release under the Freedom of Information Act.

Health and medical record data are especially in demand because of patient care needs, insurance compliance, and government regulatory requirements as well as for research (both medical and policy) and quality management purposes. Given the sensitivity of the information, it is not surprising in the face of such broad demand that there have been calls for legislation to give confidentiality protection. In the United States, the Privacy Rule of The Health Insurance Portability and Accountability Act (HIPAA)[19] began implementation in 2003. The Privacy Rule obligates most "covered entities," such as Medicare providers, to protect the confidentiality of their health care information.

Many government statistical agencies operate under multiple pieces of legislation. To illustrate this, consider the US National Center for Education Statistics (NCES). There are at least four laws that cover confidentiality of individually identifiable information collected by NCES[20]:

> *Privacy Act of 1974, as amended*—Requires Federal agencies to collect, maintain, use or disseminate any identifiable personal information in a manner that assures that such action is for necessary and lawful purpose, that the information is current and accurate for its intended use, and that adequate safeguards are provided to prevent misuse of such information.
>
> *E-Government Act of 2002, Title V, Subtitle A, Confidential Information Protection (CIP 2002)*—Individually identifiable information supplied by individuals or institutions to a federal agency for statistical purposes under the pledge of confidentiality must be kept confidential and may only be used for statistical purposes.
>
> *Education Sciences Reform Act of 2002 (ESRA 2002)*—Individually identifiable information about students, their families, and their schools remain confidential. No person may:
>
> 1. Use any individually identifiable information furnished under the provisions of this section for any purpose other than statistical purposes for which it is supplied, except in the case of terrorism (see discussion of the US Patriot Act of 2001);
> 2. Make any publication whereby the data furnished by any particular person under this section can be identified; or
> 3. Permit anyone other than the individuals authorized by the Commissioner to examine the individual reports.
>
> Individually identifiable information is immune from legal process, and shall not, without the consent of the individual concerned, be admitted as evidence or used for any purpose in any action, suit, or other judicial or administrative proceeding, except in the case of terrorism.
>
> *US Patriot Act of 2001*—Permits the Attorney General to petition a court of competent jurisdiction for an *ex parte* order requiring the Secretary of the Department of Education to provide data relevant to an authorized investigation or prosecution of an offense concerning national or international terrorism. The law states that any data obtained by the Attorney General for these purposes "...may be used consistent with such guidelines as the Attorney General, after consultation with the Secretary, shall issue to protect confidentiality."

[19] www.hhs.gov/ocr/hipaa/

[20] nces.ed.gov/statprog/2002/std4_2.asp

Beyond what is specified directly in legislation, many DSOs also must comply with regulations. For example, in the United States the Office of Management and Budget promulgated: "Order Providing for the Confidentiality of Statistical Information," Federal Register, Vol. 62, No. 124 (June 1997): 35044–55. This order establishes privacy and confidentiality policies regarding federally collected statistics for individuals and organizations. It provides language for confidentiality pledges under two conditions—first, when the data may only be used for statistical purposes; second, when the data are collected exclusively for statistical purposes, but the agency is compelled by law to disclose the data.

As the US Patriot Act of 2001 includes a legal requirement that compels NCES to share the data under the conditions specified in the law (see above), the second condition applies to NCES. In this case, the order instructs the agency to "...at the time of collection, inform the respondents from whom the information is collected that such information may be used only for statistical purposes and may not be disclosed, or used, in identifiable form for any other purpose, unless otherwise compelled by law."

On the other hand, rather than specifying confidentiality, legal requirements may actually constrain the ability of organizations to promise confidentiality. For example, letters of recommendation from a faculty member may be considered an educational record in the United States and hence under the Family Education Rights and Privacy Act of 1974 (FERPA) be accessible by a student, unless that right is explicitly waived.

1.5.4.2 Pragmatic Considerations

Even when not under legal or ethical constraints, DSOs have to be concerned about confidentiality for pragmatic reasons. Since collecting personal data is basic to the mission of DSOs, they have to be concerned with anything that might interfere with their ability to do that job.

Presumably for the pragmatic reason of maximizing the value of the research it funds, the US National Science Foundation requires data sharing, "It expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work."[21]

Confidentiality promises are also important for pragmatic reasons. Government statistical agencies, in particular, believe that respondents will provide more data and more accurate data when they have credible assurances that their responses will be kept confidential. For example, in the United States, the National Center for Education Statistics has established Statistical Standards, one of which, NCES Standard 3-2, suggests that in order to maintain high rates of response across all strata in their surveys, "privacy and confidentiality assurances citing the appropriate legislation must be provided."

[21] nsf.gov/pubs/2001/gc101/gc101rev1.pdf

There is evidence to support the belief held by statistical agencies that promises of confidentiality are useful. Research by Singer et al. (1993) suggested that promises of confidentiality lowered non-response rates and improved the quality of data provided. Singer et al. (2003) conclude that concerns about privacy and confidentiality have a small but statistically significant effect.

1.5.4.3 Ethical Obligations

DSOs take seriously ethical imperatives for confidentiality, even when no specific legislation requires it, or the pragmatic basis for providing confidentiality may by shaky. The ethical notion that an action is required because it should be done, and not just because it is compelled—say by legislation—or it is directly beneficial—say in lowering non-response rates—is certainly well established in society generally and for some time has provided much of the motivation for DSOs to afford confidentiality.

A good example of a situation where confidentiality legislation does not currently exist and pragmatic benefits to promising confidentiality—or even possible mechanisms for promising confidentiality—are not clear is with the use of high-resolution spatial data, such as satellite imagery, global positioning system (GPS) data, and output from radio frequency identification (RFID) tags for tracking people's movements. Still in this area of spatial data, confidentiality is a concern as is evident by "Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data" (National Research Council, 2007) which argues for attention to confidentiality and data sharing on the basis of professional codes of ethics and, more pragmatically, any pledges that may have been made to the agencies funding the research. In its code of ethics adopted in 2003, The Urban and Regional Information Systems Association, for example, exhorts its members to "Hold information confidential unless authorized to release it."[22]

Are there some general ethical principles that should guide data stewardship? The United States, the European Union, and a growing list of countries embrace pluralism, public decision making based on representative democracy, and substantially a market-oriented economy. Broadly, in such a society a variety of contrasting and competing individuals and interest groups vie for influence and benefit. To have this process work properly requires extensive generation of information about persons, and its dissemination to a variety of actors. Consider the issue of primary education. Parents in making choices for their children want information about a school's effectiveness. Government in developing public policy decisions about education wants to know what programs work and what programs do not. Advocacy organizations want data to support their positions. Yet this evident demand for data must be tempered by other considerations, considerations that constrain each aspect of the CSID data process.

[22] urisa.org/about/ethics

Actual protection of the confidentiality of microdata is a function and an obligation of all DSOs even as it might vary, depending on their mandate, political and cultural milieu, and legal auspices. To see this commonality, compare the perspective on confidentiality of the Office for National Statistics (ONS), which provides official statistics for the United Kingdom, to that of the US Census Bureau, which is the largest statistical agency in the United States.

Both the ONS and the Census Bureau have official statistical responsibilities to their nation. ONS is the executive office of the UK Statistics Authority, a non-ministerial department which reports directly to Parliament.[23] A part of the Department of Commerce, the Census Bureau, produced in 2003 the document "U.S. Census Bureau Strategic Plan FY 2004–2008"[24] (accessed January 7, 2006), which detailed its thinking. We summarize the perspectives of the ONS and the Census Bureau in Table 1.1, drawing some rough parallels.

1.6 High-Quality Statistical Data Raise Confidentiality Concerns

A charge to national statistical offices, and other DSOs, is to provide data users with access to high-quality statistics. Official statistics supply the factual information necessary to guide public policy through a variety of empirical appraisals, notably estimating population changes, assessing unemployment, determining health care needs, and measuring macroeconomic factors. The quality of official statistics depends on technical factors such as sample selection methodology, treatment of missing data, appropriate measures of variability and bias, and methods of comparison across population subgroups and across countries. Quality also depends critically on ensuring acceptable response rates and the accuracy with which respondents provide data. Complicating this charge to DSOs to provide high-quality statistical data is the fact that the characteristics that make statistical data of high quality also make confidentiality of real concern. We now explore why this is the case.

1.6.1 Characteristics of High-Quality Statistical Data

Certainly, to be of high quality all statistical data sets should be accurate, accessible, and comprehensive. In many cases these data sets should also contain records that are individual, geographically specific, hierarchical, and longitudinal. Incongruously, these are the very characteristics that pose problems for confidentiality protection. As we will see, this is an unfortunate and pervasive feature of the statistical confidentiality problem. Exactly the same data structures which add value

[23] http://www.ons.gov.uk/about/who-we-are/index.html

[24] http://www.census.gov/main/www/strategicplan/strategicplan.html

Table 1.1 Expressed responsibilities of the lead statistical agencies in the United Kingdom and the United States

Attribute	United Kingdom's Office for National Statistics	United Status Census Bureau
General mission	An independent, fact-based, voice on matters of social and economic policy that commands the trust and confidence of the public.	Support the economic and political foundations of the United States by producing benchmark measures of the economy and population for the administration and equitable funding of federal, state, and local programs.
Facilitating governmental accountability	A mechanism of accountability of the government to the people.	Meet constitutional and legislative mandates by implementing a reengineered 2010 census that is cost effective, provides more timely data, improves coverage accuracy, and reduces operational risk.
Data users	A source of high-quality statistical information to the Government, Parliament, and the wider community for issues of public debate. A statistical service to business which promotes the efficient functioning of commerce and industry.	Meet the needs of policy makers, businesses and nonprofit organizations, and the public for current measures of the US population, economy, and governments.
Support for researchers	A statistical service to researchers, analysts, and other customers.	*Not mentioned*
International responsibilities	A statistical service to meet European Union and international requirements.	*Not mentioned*
Confidentiality	Protective of the confidentiality of people and organizations from the unauthorized disclosure of information held about them by their departments for National Statistics or other statistical purposes.	Support innovation, promote data use, minimize respondent burden, respect individual privacy, and ensure confidentiality.

to a data set for a user also add disclosure risk. Let us address each characteristic in turn to see first why the characteristic is appropriate for quality data and then see for each characteristic why it complicates the assurance of confidentiality.

- *Accurate.* While this may seem an evident requirement, it is complex conceptually and a challenging task to achieve in practice. Values for a particular attribute for a particular population unit, such as total income last year for a household in

Adelaide, Australia, should be assessed with low variability and minimum bias. Suppose total income is to be assessed through a personal interview survey. Who from the household qualifies as a respondent? How should “total income” be explained? Is there a recall bias, where respondents consistently under or over assess what has happened to them? Values across attributes may be incompatible, as when a respondent to a questionnaire says on one question that they had no pets when a hurricane struck, but on a later question says that they chose not to evacuate because of concern for their pets. Values across entities may possibly be incompatible, as when a reported total income for one household is ten times higher than the next highest reported for similar households. Dealing with such outliers presents serious difficulties for a data analyst. Further, a serious problem in data gathering is how to curtail both erroneous response and non-response. Erroneous response creates obvious problems for the data analyst, but non-response is also vexing to a data analyst when the lack of information from the non-respondents may selectively bias the reported values.

- *Accessible.* The quality of data products can hardly be manifest unless analysts can first gain access to the data. Although the most sophisticated data users are typically willing to expend much effort in seeking appropriate data for their studies, there are limits on the time and resources they can devote to this quest, indeed, limits on how much frustration they will tolerate. Today, researchers and other data users have become accustomed to obtaining easy electronic access over the web.
- *Comprehensive.* Certainly, comprehensive data must fully encompass the range of features that interest analysts, and so must have attributes that are relevant to the questions that need to be addressed. Generally, since analysts will want to make comparisons and draw out relationships, the data must be multivariate. In a community health study, for example, many variables could well be measured jointly for each of the subjects. It could address a wide variety of health conditions, such as, heart disease, asthma, sleep disorders, learning disabilities, obesity, and diabetes. The study could relate these variables to risk factors, such as nutrition, activity levels, smoking, blood pressure, social and economic disparity, and health care access, whilst also comparing them with other communities. In the United States, for example, in a health study of Hispanic communities, researchers could compare a majority Mexican-American community with a majority Cuban community, a majority Puerto Rican community, and a majority Central-South American community.
- *Individual.* To do their job, knowledgeable researchers often want individual-level rather than aggregated data. Thus they want access to *microdata*, that is, the records on individual study participants. This allows them to build models for relationships that could not be constructed from simple tables of counts or summary statistics. In the community health study, they might then be able to determine any relationship between sleep disorders and blood pressure using regression models.
- *Hierarchical.* Many researchers need hierarchical information to answer specific research questions. For example, to answer questions about work–life balance

or the division of unpaid labor in the home or the relationship between household structure and individual social capital, researchers need information about households plus individual-level information about members within a household. Similarly, linked employer–employee data are invaluable for studies of employment.

- *Geographically specific.* For many social science and public policy studies, place is a critical attribute. An empirical study of youth gang activities, for example, would be vacuous without considering in some detail the "turf" of rival gangs. Similarly, local policy makers want data that are specific to their area.
- *Longitudinal.* Often analysts want to observe changes over time, just as a study of health and nutrition of preschool children in four Guatemalan villages[25] followed up on them for a period of 5 years to see what changes had taken place in their health and life circumstances. Without such longitudinal data, which tracks individual subjects over time, we have no basis for attributing observed changes to any causal mechanism.

1.6.2 Disclosure Risk Problems Stemming from Characteristics of High-Quality Statistical Data

With the evident appeal of these characteristics, it is unfortunate that it is just such data that have the highest risk of disclosure and so present the most problems for confidentiality protection. Now, consider each characteristic from the perspective of the concerned data provider or privacy advocate.

- *Accurate.* The confidentiality problem is that accurate values may make it possible to link data records to an identified database. To illustrate how this can happen, in Allegheny County, Pennsylvania, real property records are publicly available on the web and are identified with the owner of the property. For residences, the record gives the size of the lot, say 7504 square feet. A housing survey might ask for the square footage of the owner's lot. If both the survey and the property tax database had this figure accurately at 7504 and there was only one such property in the county, it would be possible to link the name of the owner with the survey record. This would reveal for that owner all other attribute values; for example, race, income, and sexual orientation that might well have been collected in the housing survey.
- *Accessible.* An obvious way of protecting confidentiality is by making access difficult. Indeed, often access is made difficult through a series of administrative hurdles established in the cause of confidentiality protection that a would-be data user must clear. Examples are restricting who may access the data, where they may access it, limiting the analyses they may conduct, and only allowing access to a subsample of the data.

[25] http://www.unu.edu/unupress/food2/UIN09E/uin09e04.htm

- *Comprehensive.* Generally comprehensive data have a wide range of attributes. This provides more opportunities for record linkage with identified databases because there are more chances for an overlap of attributes. Note that linkage in the housing example depended on lot size being assessed in the survey. Furthermore, data for broader surveys typically have more attributes that would be considered sensitive if they were revealed, like race, religion, income, and sexual orientation.
- *Individual.* Certainly it is individual-level, rather than aggregated, data that directly provide information on an individual, and so have the potential for identity disclosure.
- *Hierarchical.* Hierarchical household information increases identification disclosure risk particularly for large households.
- *Geographically specific.* Obviously the more detail on where someone is (lives, works, plays), the easier it is to identify them.
- *Longitudinal.* A survey record that indicates that the respondent worked as an electrical engineer in Madrid in 1999 places them among thousands, as would a similar record that indicates they worked as an electrical engineer in London in 2006. Linking the two together in a longitudinal data set may make them unique, and so more easily identifiable.

1.7 Disclosure Risk and the Concept of the Data Snooper

As we have seen DSOs are appropriately worried about the risk of a disclosure. To make this risk concrete, DSOs often speak about an attack by an illegitimate user of the data. Who might want to attack statistical confidentiality? There is little hard evidence of anyone actually attempting to use statistical data products to compromise confidentiality.

Because many DSOs, and most especially statistical agencies, view the stakes in terms of their reputation, their future ability to collect data, their professional responsibilities, and a variety of legal requirements to be so high, they find it useful to talk of those who might mount an inferential attack as a single, almost cartoon-like, character. A popular name for this character is the *data snooper*, but various names have been applied, such as, data spy (seems more malicious than can be justified), data intruder (our character wants to see individual-level information not just to trespass), and not the attacker (our character does not seek to destroy but just to get information, albeit unauthorized information) (Duncan and Lambert, 1989; Duncan, 2003; Heitzig, 2005).

Writers usually find that it eases discussion of confidentiality and disclosure limitation to use language like "make life difficult for snoopers" (Domingo-Ferrer et al., 2005). Of course, a potential attacker need not be a single individual. Certainly an attacker could be a government agency, a corporation, or an interest group. Nor, given the stakes, does there have to be compelling evidence of the existence of a data snooper for DSOs to want to ensure that they have appropriate security in place to prevent a confidentiality attack.

Elliot and Dale (1999) and Paass (1988) explore the psyche and motivations of the data snooper. The snooper may or may not be someone with limited access to the data and may or may not be motivated for malicious reasons. Prudently, however, the statistical agency must assume that the data snooper has access to sophisticated analytical tools, is knowledgeable about the data and has ready access to relevant external data sources, and has the necessary computational capability to attempt an attack on the data. Based on this understanding, the DSO requires effective SDL methods, like those described in Chapters 4 and 5.

What might the data snooper do? Rasinski and Wright (2000) say that if the data snooper "learns the identity of a survey respondent the snooper may use the survey data to embarrass, harass, or harm the respondent." Having admitted that this is a real and present risk, how can the DSO assure itself that adequate steps have been taken to protect against an attack by a data snooper? What opportunities does the DSO have to ensure that it can continue its mission of providing quality data while protecting confidentiality? These are key questions that we now begin to address.

1.8 Strategies of Statistical Disclosure Limitation

To fulfill its dual mandate of providing statistical data while protecting confidentiality, the DSO can employ strategies that come in two flavors: restricted data and restricted access (see Duncan and Pearson (1991) and Little (1993)). The key idea is to apply procedures that will encourage a data provider to trust that the DSO can provide confidentiality.

The confidentiality of individual information can be protected, thus limiting disclosure risk, by restricting the amount of information provided or by adjusting the data in released tables and microdata files (*restricted data*) or by imposing conditions on access to the data products (*restricted access*), or by some combination of these.

1.8.1 Restricted Access

There are three primary procedures that DSOs can use to provide restricted access to confidential data: Research Data Centers (RDCs), Restricted Statistical Software, and Licensing Agreements. RDCs permit use of confidential files in a physically secure environment. Data users agree to abide by specified conditions governing the access and use of the confidential data. Products of their research are reviewed by the DSO staff to assure no confidential information is revealed. Under *Restricted Statistical Software* a data user is required to access the data only through specialized software, often remotely over secure electronic lines to dedicated computers. The statistical products are reviewed by the agency to assure no confidential data are revealed. Under *Licensing Agreements* a researcher is permitted to use confidential data offsite, but only under restricted conditions as specified in a legally binding agreement. Typically these agreements subject the user to fines and/or being denied access in the future for violations of the stipulated conditions of use. In many cases users are subject to external audits conducted by the agency to ensure terms of the agreement are being followed.

Readers interested in restricted access can find detailed information in Jabine (1993a) and Duncan et al. (1993), *Private Lives and Public Policies*, p. 157, and "Restricted Access Procedures" by the Confidentiality and Data Access Committee (April 2002).[26] See also concerns raised in Fienberg (1993). In this book the topic of restricted access will be continued in Chapter 7.

1.8.2 Restricted Data

Restricted data result from transformations of source data. The transformations are designed to lower disclosure risk to an acceptable level while maintaining the utility of the resulting data for statistical purposes (see Dalenius, 1988; Duncan and Lambert, 1986; Bethlehem et al., 1990; and Jabine, 1993b). Chapter 4 examines restricted data procedures for tabular data and Chapter 5 examines restricted data procedures for microdata. While early work was done primarily at the US Census Bureau (see, e.g., Barabba and Kaplan, 1975 and Greenberg, 1990), research programs for developing techniques for statistical disclosure limitation and producing restricted data accelerated in the 1990s. In 1992, Eurostat and the ISI (International Statistical Institute) organized the International Seminar on Statistical Confidentiality in Dublin, Ireland. In 1992, a special issue of Statistica Neerlandica (Vol. 46, No. 1) was dedicated to disclosure limitation. In 1993, a special issue of the Journal of Official Statistics, Vol. 9, No. 2, was dedicated to disclosure limitation. That issue contains the papers which were presented at a workshop sponsored by the Panel on Confidentiality and Data Access, of the Committee on National Statistics of the National Academy of Sciences. The panel report is entitled *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, by Duncan et al. (1993), and was published by the Committee on National Statistics and the Social Science Research Council, National Academy Press, Washington, DC.

1.9 Summary

Our goals in this chapter have been to answer several key questions beginning with the basic one of just what is statistical confidentiality. In Section 1.1 we first defined the broad concept of confidentiality this way: "under confidentiality, the party holding the information is bound by an implicit or explicit promise that it be protected from unauthorized or inappropriate access and usage." We then laid out our conception of statistical confidentiality as a body of principles, concepts, and procedures

[26] www.fcsm.gov/committees/cdac/cdacra9.doc

that permit confidentiality to be afforded to data, while still permitting its use for statistical purposes.

In Section 1.2 we identified the stakeholders in the statistical process and what their stakes are by discussing a medical data example.

We then laid out in Section 1.3 the fundamental dilemma in statistical confidentiality, which is that the two basic missions of a DSO, to provide access to statistical data and to maintain confidentiality, are fundamentally incompatible, yet both missions are required. Resolving this dilemma is the challenge for statistical confidentiality.

In Section 1.4 we answered the question of why statistical data are in such demand by giving real-world examples of how policy analysts and researchers need statistical data to address problems of public interest such as providing affordable housing.

In Section 1.5 we answered the question of why DSOs are so concerned about confidentiality. Certainly from the 4th century BC. with the Hippocratic Oath there has been a concern for confidentiality. Today, a combination of societal factors and technological changes has raised the demand for personal information. In education, for example, we now need life-long learning to be provided wherever we may be in the world. Continuous assessment of progress and storage of results in electronic databases facilitates learning, promotes accountability, and assures employers of qualifications. As noted in Section 1.4, the dilemma a DSO faces is that there is increased public concern for confidentiality, to the extent that the statistical enterprise of providing data access is threatened. We explored four interrelated reasons that the statistical community is in a crisis about data access: privacy worries, confidentiality concerns, a changing legal and social context, and an increased sensitivity to social impact, all of which relate to whether a data provider can trust a DSO. We examined the three fundamental motivations for DSOs to develop policies and methods to protect confidentiality: a confidentiality breach may violate ethical norms, it may compromise the ability of a DSO to do its job (which is a pragmatic consideration), and it may be required by law.

Yet, as we discussed in Section 1.6, high-quality statistical data raise particular confidentiality concerns. All statistical data should be accurate, accessible, and comprehensive. In many cases they should also be individual, geographically specific, hierarchical, and longitudinal. As we show in this section these are the very characteristics that pose problems for confidentiality protection. Exactly the same data structures which add value to a data set for a user also add disclosure risk.

Our topic in Section 1.7 was disclosure risk and who is the data snooper who might want to attack statistical confidentiality. Although there is little hard evidence of anyone actually attempting to use statistical data products to compromise confidentiality, most statistical agencies view the stakes in terms of their reputation, their future ability to collect data, their professional responsibilities, and a variety of legal requirements to be so high that they must provide thorough confidentiality protection. They find it useful to talk of one who might try to compromise confidentiality as a data snooper. The data snooper could do this if the disclosure risk of a released data product were too high.

Finally in Section 1.8, we began to address the big question of how statistical confidentiality can be protected. The confidentiality of individual information can be protected, so limiting disclosure risk, by restricting the amount of information provided or by adjusting the data in released tables and microdata files (*restricted data*) or by imposing conditions on access to the data products (*restricted access*) or by some combination of these.

The remaining chapters of this book build on the answers to these foundational questions. Beginning with Chapter 2, we lay out the basic concepts of statistical disclosure limitation, the key tool for protecting statistical confidentiality.

Reidentification of Chapter Quotation:

Hubert H. Humphrey, Vice-President of US 1965–1969; Democratic presidential candidate 1968 (1911 – 1978)

Original quotation: “It was once said that the moral test of Government is how that Government treats those who are in the dawn of life, the children; those who are in the twilight of life, the elderly; and those who are in the shadows of life, the sick, the needy and the handicapped.”

Chapter 2
Concepts of Statistical Disclosure Limitation

The SDL literature has its own terminology. Understanding this terminology and, more importantly, the concepts underlying the terminology is essential to learning how statistical confidentiality can be best employed. In this chapter we look at the structure of disclosure risk, its assessment, and its limitation. Complicating our task is that many terms, such as "protecting data" or "sensitive data," have no universally accepted meaning. Driven by variations in their historical and legal environment, DSOs exhibit differences in how they use terminology. This can lead to confusion in discussions among DSOs and indeed within the SDL research community as well. In this chapter we lay out widely accepted concepts and a terminology that provides a common framework intended to minimize confusion. These concepts and terminology are used consistently throughout the rest of this book.

2.1 Conceptual Models of Disclosure Risk

DSOs develop confidentiality policies and procedures in order to address the concern that the data products they disseminate may disclose the identity of a respondent or information about the characteristics of a particular respondent. For a DSO to assess disclosure risk, it must understand what is being risked and what are the consequences of a disclosure.

DSOs need to preserve the trust accorded to them by respondents. Assurances of confidentiality appear at the top of any responsible survey questionnaire or are provided at the start of an interview. Because of these confidentiality statements, the need to assure data protection becomes explicitly part of the data processes from collection through to dissemination. There is a social contract between the DSO and the respondent that the respondent's information will be looked after and any data dissemination arising from those responses will be carried out responsibly.

DSOs often treat disclosure as if it would be a catastrophe, doing major damage to their mission. Many DSOs, especially National Statistical Offices, believe that if a disclosure happens it will have consequences far beyond the informational impact of the disclosure for the individual for whom disclosive attributions have been made. They believe that a single publicized disclosure could lead to such a breakdown in trust that there will be a significant reduction in respondent cooperation. A

G.T. Duncan et al., *Statistical Confidentiality*, Statistics for Social
and Behavioral Sciences, DOI 10.1007/978-1-4419-7802-8_2,

breakdown of trust leading to non-cooperation is a sociological construct that is supported by historical evidence. In a review of the census experience of 11 western European countries, McDonald (1984) found significant concerns for privacy and confidentiality in England, the Federal Republic of Germany, Italy, the Netherlands, Norway, and Sweden. In particular, the German census originally scheduled for 1981 was postponed for budgetary reasons, but then in 1983 there were large-scale public protests and over a thousand lawsuits were filed against the proposed census. Many reasons could be advanced for this public agitation including worries about increased use of microdata, such as the specter of advanced computer technology, and lack of support for the value of the census by the media and government officials. In addition to these possibilities, Butz (1985) posited the reason of fears that government agencies would use census data against individuals. Presumably, these fears were particularly salient in Germany given the abuses of census data under the Nazi regime. In March 1983, the Constitutional Court suspended the census. The Court argued that the planned dissemination procedures lacked adequate confidentiality protection (Butz and Scarr, 1987). The census was not held until 1987.

However, the first step in the sociological reasoning that a single publicized breach would lead to a reduction of trust is less justified. It is by no means clear that a reduction in trust necessarily follows from a single breach nor that it is unmediated by the DSO's reaction to that breach (Mackey and Elliot, 2009). The paradox is that there are only isolated cases publicly known of a statistical disclosure event having happened. Therefore, we have no real evidence either for or against the proposition. In the face of such lack of evidence, DSOs feel obliged to maintain the view that such events would be near catastrophic and therefore view only minimal disclosure risk as acceptable.

To further complicate this already fuzzy picture, it is clear that zero (or even effectively zero) disclosure risk is not compatible with a DSO meeting its obligation to disseminate data that have high utility. In their actual behavior DSOs must therefore be more pragmatic. In the United Kingdom, for example, the policy of the Office for National Statistics (ONS) is that no disclosure takes place if more than a "reasonable amount of effort" on the part of a data snooper would be required to break confidentiality. Underlying this pragmatism is the interplay between perceived and objective risk. If the risk is perceived as negligible then would-be snoopers are likely to look for other ways to achieve their goals. Importantly therefore a reduction in a data snooper's perception of disclosure risk is *de facto* a reduction in the objective risk.[1]

Marsh et al. (1991) used conditional probabilities to formalize this interplay as follows:

$$P(\text{identification}) = P(\text{identification}|\text{attempt}) \cdot P(\text{attempt})$$
$$P(\text{identification}|\text{no attempt}) = 0$$

[1] This is analogous to a fake burglar alarm box.

We can limit the actual risk by controlling either or both the probability of an attempt being made as well as the probability that it is successful given that it has been made. Undoubtedly, DSOs engage in both forms of control, sometimes simultaneously. For example, the United Kingdom's ONS used a record swapping algorithm for its 2001 census outputs. A proportion of records between 0 and 5% had their local geographical codes swapped. ONS announced this range for the swap rate, thus having some impact on the disclosure risk. Within the announced range the swap rate could be zero, so the impact on objective risk would also be zero. But whatever the actual swap rate—provided it is keep secret, a reduction in risk is achieved beyond the objective impact of the actual perturbation. The reason for this is the increase in the uncertainty for any inferences that a data snooper might make. This shows that effective public communication is as important in disclosure risk management as is manipulation of data before dissemination.

2.1.1 Elements of the Disclosure Risk Problem

The usual reasoning about disclosure risk is this: a data snooper who is in possession of some information about one or more identified population units wants to infer further information about those population units. The data snooper seeks to do this by linking that information to items in the target data set. The main alternative to this line of reasoning is the spontaneous recognition of a data element by a bona fide user of the data.

This basic formulation has different meanings depending on whether we are considering aggregate data or microdata. Chapter 4 provides a detailed development for aggregate data; Chapter 5 does the same for microdata. Here we introduce the basic concepts for each.

2.1.1.1 Microdata

Microdata are the raw material at DSOs. They are the direct information collected from the respondents. Traditionally microdata are organized in a database, where each record is an encoding of the answers of each contributor (person, household, organization, etc.) to a survey or census. In principle, though, microdata could be collected in a variety of ways. For example, economic microdata at a country-level could be generated from a study of national accounts; hospital in-patient data could be collected from hospital administration records; and data on how the response rate and accuracy of a census might change with alterations in the number of questions presented could be collected through an experiment. The crucial point about microdata is that for each population unit represented in the data set there is a record of the values of multiple attributes.

A microdata set that contains all the variables collected for the whole population is called the *full table*, acknowledging that microdata are transformable into frequency tables. Invariably, the DSO stops short of releasing the full table. Instead, it may release subsets that are marginal tables of the full table or possibly samples of microdata (or indeed it may have only collected a sample as in a social survey). If it does release a sample of microdata then it will usually reduce the number of variables that it releases as well as the level of detail on those variables it does release (particularly temporal and geographical detail). Limiting the data to be released is typically the initial phase of statistical disclosure limitation. Before looking at SDL, we examine the event we are trying to limit—"disclosure."

In the case of microdata there are two different concepts related to a disclosure: *identification* and *attribution*. Identification is the association of a known population unit with a particular microdata record. Attribution is the association of information in a microdata set with a particular population unit (see Lambert, 1993; Reiter, 2005b).

Operationally, the data snooper must usually identify a population unit in order to make attributions about it. It is the attribution itself which forms the disclosure. However, there are cases where identification and attribution occur independently. Identification may occur without attribution if the data snooper already knows all the information contained within the microdata set about the population unit in question, so no new information was attributed to the population unit from the linkage (an obvious example of this being self-identification). Conversely, where two or more population units with a given set of variable values necessarily share another attribute, the attribution of that additional information to each population unit for which the key information is known can take place even without direct identification. This could arise, for example, where a data set contained occupational and income information and it was known that all individuals with a given occupation had a particular income. Notwithstanding these two cases, the canonical process is identification leading automatically to attribution.

2.1.1.2 Deliberate Linkage

Deliberate record linkage (or *matching)* is the typical mode of disclosure. The presupposition is that a data snooper has access to a data set which contains direct identifiers for population units (name, address, etc.) and a set of *key variables* which are also present in the target data set. The key variables are then used to link the identifiers to the data snooper's target, as shown in Fig. 2.1.

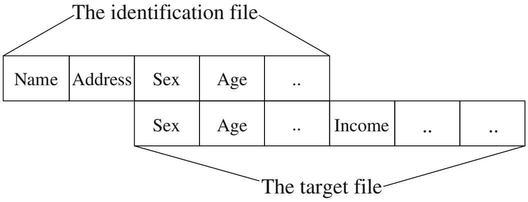


Fig. 2.1 An illustration of key variable matching leading to disclosure (after Elliot, 2000)

2.1.1.3 Aggregate Data

Aggregate data are most commonly published as tables, often tables of counts (frequency tables), but also cross-classified tables of statistics such as means and sums. Tables of counts are aggregate data obtained from the microdata. Characteristic of a statistical table is that it also contains marginal sums, so that in addition to the data values, we also have relations linking these values, as for a sex variable a marginal total may be the sum of the number of males and the number of females. These links between values add complexity to the problem of protecting tables which does not exist when protecting microdata.

With tabular data there are two conceptual processes underpinning the notion of disclosure—*subtraction* and *attribution.* Aggregate data are often data for the whole of a population (where this is not the case the aggregate data are effectively sample microdata for our purposes). Attribution with aggregate data is similar conceptually to that with microdata; that is, it is the association of information in an aggregate data release with a particular population unit. Importantly however, whereas with microdata, attribution follows automatically from correct identification, with aggregate data, attribution follows contingently from the presence of non-structural zeroes anywhere in the aggregate data set.

2.1.1.4 Attribution and Subtractive Attack

Subtraction is the removal from an aggregate data set (which could be a single table or set of tables) of particular individuals for which the values or all the variables within the aggregate data set are known to the data snooper. To clarify attribution and subtractive attack for tabular data consider Table 2.1. Take the population represented in this table to be everyone at a workshop you are attending. Over coffee, you overhear someone saying that they earned over 1 million pounds in the last quarter. Given that you know the data in Table 2.1, you can infer that person is a lawyer. This is positive attribution—the association of a particular value with a particular population unit. Conversely, if you talk to some people and find out that they are academics, you can infer that they do not have a high income. This is a negative attribution—the disassociation of a particular value for a variable from a particular population unit.

Table 2.1 Table of counts of income levels for two occupations from hypothetical population

	High	Medium	Low	Total
Academics	0	100	50	150
Lawyers	100	50	5	155
Total	100	150	55	305

Note that association and disassociation are different forms of the same process, attribution arising from zeroes in the data set. The presence of a non-structural zero in the internal cells of a table is potentially disclosive. In effect, the attributive process is disassociative.

Now consider Table 2.2. The population in this table differs in one respect—there is one highly paid academic. Give this table, we can no longer make the inferences that we could from Table 2.1 (at least not with certainty). However, what about myself? I am a member of the population represented and I know my own occupation and income. Suppose that I am a highly paid academic. Given this extra-table information, I can subtract 1 from the high-income cross academic cell in the table, which then reverts to Table 2.1. I am then back to the situation where I can make disclosive inferences from overhearing partial information about particular individuals.

Table 2.2 Table of counts of income levels for two occupations from hypothetical population

	High	Medium	Low	Total
Academics	1	100	50	151
Lawyers	100	50	5	155
Total	101	150	55	305

We can extrapolate this further. Consider a situation where you have complete information (for the two variables contained in Table 2.2) about multiple individuals. In effect such information represents a table of counts of the subpopulation of the individuals for whom I have complete knowledge. On the assumption that identification is available for both that subpopulation and for any additional information you gain through overheard conversations (or other sources), you can subtract the whole of that table from the population table before proceeding. In principle, this could lead to more zeroes appearing in the residual table. This is illustrated in Tables 2.3 and 2.4. The "low-paid lawyers" cell would be particularly vulnerable to further subtraction and this illustrates a further crucial point: whilst zero counts are inherently disclosive, low counts also represent high disclosure risk, in that,

Table 2.3 Table of counts of income levels for two occupations for a subpopulation of Table 2.2 about which a hypothetical snooper has complete knowledge

	High	Medium	Low	Total
Academics	0	80	25	105
Lawyers	70	20	5	95
Total	70	100	30	200

Table 2.4 Residual table of counts resulting from subtracting Table 2.3 from Table 2.4

	High	Medium	Low	Total
Academics	1	20	25	46
Lawyers	30	30	0	65
Total	31	50	30	111

with low cell counts, it will be easier to obtain sufficient information external to the aggregate table to enable subtraction to zero than with high cell counts.

Beyond the subtraction to zeroes there is another sense in which low cell counts constitute a risk. Consider again Table 2.2. Recall that without external information about the population represented in the table it is impossible to make inferences *with certainty* about an individual given partial information about that individual. However, imagine again that you overhear someone at the workshop boasting about their high income. Although not sure that this individual is a lawyer, you can assert it so with a high degree of confidence. From the table, the conditional probability that a randomly selected individual is a lawyer, given that they are higher earner, is greater than 0.99. Depending on my purposes in disclosing this information, this may be good enough to do so.

The principle of subtraction is also central to understanding attacks on magnitude tables. Consider Table 2.5 with hypothetical total consultancy earnings by statisticians in the University sector in a certain year.

Table 2.5 Hypothetical magnitude table

	Consultancy sales (Euros)
United Kingdom	700,000
Spain	1,000,000
France	2,000,000
Germany	380,000
Italy	700,000
Netherlands	460,000

Let us suppose that I am in charge of Consultancy sales at Camford University in the United Kingdom and know thereby that we earned a total of 550,000 Euros last year. I therefore know, by subtraction, that all of the other UK universities combined have earned 150,000 Euros. This illustrates the principle of dominance. As my university dominates the cell for the United Kingdom, I (or anyone else who has this information about my university) am able to make some limiting inferences about other contributants to the cell.

2.1.1.5 Linking Tables

A further problem that arises with aggregate data is that of table linkage. This occurs when two or more tables in a release of aggregate data have variables in common (they are said to overlap). In effect, all aggregate tables of counts are margins of the *full table* (the underlying microdata), and from which, given enough external information, a user of the aggregate data could reconstruct that full table. Even though this is unlikely, a more pragmatic concern is that a data snooper might combine tables by linking subtables to recover larger tables and these larger tables themselves may be disclosive, or more vulnerable to a subtractive attack. For example, consider Tables 2.6–2.8, all drawn from the same hypothetical population.

Tables 2.6–2.9 An illustration of how groups of tables might be disclosive

Table 2.6

	Var 1	
Var 2	A	B
C	3	9
D	2	2

Table 2.7

	Var 1	
	A	B
E	1	10
F	4	1

Table 2.8

	Var 2	
Var 3	C	D
E	8	3
F	4	1

Table 2.9

	Var 1 and 2			
Var 3	A,C	A,D	B,C	B,D
E	0	1	8	2
F	3	1	1	0

None of these three tables is themselves disclosive, although we note some low counts and therefore we consider the cells to be risky.

However, because these three tables overlap and form the margins of a three-way table, we can, through linear programming or other means, identify all of the possible three-way tables that correspond to the two-way margins. In this case there exists only a single feasible three-way table (Table 2.9) that has the two-way margins shown in Tables 2.6–2.8, and in that three-way table there are two cells with zero counts. Since this table contains zeroes, it is potentially disclosive and therefore so is the release of Tables 2.6–2.8, even though those tables contain no zeroes. For extensions of these ideas see Chowdhury et al. (1999).

The situation is complicated further when we consider table linkage in interaction with the subtraction problem discussed above. Even if a set of tables are not linkable to form a single feasible larger table, subtraction of individual population units for which complete information for the larger table is known might lead to a single feasible table. For example, adding an individual with characteristics AED to the data underlying Tables 2.6–2.8 leads to more than one feasible table, so removing that individual leads to the reduction of the number of feasible three-way tables to 1.

Clearly, risk assessment needs to take account of all of the above possibilities. In Chapters 3 and 4 we will discuss methods for doing so.

2.1.1.6 Hierarchical Tables

A table is said to be a "hierarchical table" when one of the variables is a hierarchical variable. This means that there exists a variable which assumes several values, each one decomposable into other values, and so on. An example of a hierarchical variable is "geographical location," which may represent countries, regions, counties, local authorities, and so on. The existence of hierarchical variables in a table creates additional links (aggregations) that complicate protecting confidential data.

2.1.1.7 Linking Anonymized Data Sets

As well as the internal linkages that we have discussed, data from different data sets might be linked. In fact there is now a well-established field of research concerned with how to link anonymized data sets legitimately, especially to improve data quality or the ability to improve statistical inferences, and so on (see for example Tranmer et al., 2005). However, this kind of linkage also has an impact on the risk of disclosure; Smith and Elliot (2004, 2005) show that it is possible to improve the ability to link two distinct samples of microdata if an aggregate table of population counts for a subset of the microdata is introduced into the mix. So having multiple data sets available in the same data environment adds an extra layer of complexity onto assessing disclosure risk.

2.1.1.8 Spontaneous Recognition

The notion of spontaneous recognition is simple. You know of person X who has an unusual combination of attribute values. You are working on a data set and observe that a record within that data set also has those same attribute values. You infer that the record must be that of person X. In order to be truly spontaneous you must have no intent to identify. Otherwise this is just a specific form of deliberate linkage. Note that you could, of course, be wrong in this inference. Even if the record was unique, the data set may only be a subset (as in a sample survey) of the population and the record may not be unique in the population (Skinner and Holmes, 1992; Bethlehem et al., 1990).

2.1.2 Perceived and Actual Risk

So far we have focused on the "objective" components of the risk based on the data to be released and its relation to what a data snooper may or may not know. However, intrusion also has a subjective component which can have an impact on overall risk. Here are examples:

(1) The perceived sensitivity of the target data not only affects the motivation of a data snooper but also affects the likelihood of non-response.
(2) The perceived likelihood of success of an attack affects the motivation of data snoopers.
(3) The presence of low counts in aggregate data may make respondents believe that the data are risky, independently of whether they are or not.
(4) Common sense demographic knowledge may frame individual's perception of unusual combinations of characteristics in microdata (this is the essence of the spontaneous recognition situation).

The exact relationship between perceived and actual risk is complex. Clearly, with Example 2 above there is a feedback loop whereby the subjective and objective elements of risk are causally related. This is another reason why measuring disclosure risk in an absolute sense is difficult, if not impossible, and why SDL researchers tend to focus on the objective component. Nevertheless, pragmatically speaking, respondent co-operation depends directly on the subjective component and only via this on the objective component. The subjective component of disclosure risk is therefore important and DSOs should pay at least as much attention to managing that perception as to directly controlling the actual risk.

2.1.3 Scenarios of Disclosure

Once we understand that disclosure risk management is as much about psychology as about objective characteristics of data, we naturally turn to thinking about how a disclosure event might occur. Who are data snoopers? What are they trying to achieve by their snooping? Answering such questions is an important first step in risk management. Elliot (2000, 2005) and Elliot and Dale (1999) have produced a classification for the analysis of snooping attempts that enables the generation of *key variables* available to a data snooper and therefore should be considered in any risk assessment:

- Motivation
- Means
- Opportunity
- Attack types
- Key/matching variables
- Target variables
- Effect of data divergence
- Likelihood of Success
- Goals achievable by other means?
- Consequences of attempt
- Likelihood of attempt
- Effect of variations in database structure

Let us consider each of these items in turn.

2.1.3.1 Motivation

The motivation for a disclosure attempt comprises two elements:

Rationale: A description of the motivations of the data snooper (for example, to discredit the DSO).

Goal: A specification of the state of the world that the data snooper wishes to achieve; that is, an operational definition of what would be achieved by the disclosure attempt (for example, the release of a match into the public domain).

2.1.3.2 Means

There are three elements in how an attack would be made:

Skills: The statistical and computational skills that would be needed to select and apply an adequate matching technique and to interpret the results.

Knowledge: Any factual knowledge that assists the data snooper in their attempt. As well as information stored in databases, available knowledge might include knowledge about a particular locality (such as the prevailing housing type) and direct knowledge (e.g., by visiting an address you can establish the housing type or by talking to an individual you can establish their occupation).

Computational power: Enough computing power to perform the analyses required to achieve the goals of the data snooper. Because of readily available computer hardware and inexpensive mathematical software it makes sense to assume that all data snoopers can compute (for example) the maximum and the minimum values for any missing data in a table. Today a prudent DSO takes a worst-case attack using high-performance computing as the expected attack.

2.1.3.3 Opportunity

Without access to the data set the snooper lacks the opportunity to break confidentiality. Access may be limited because the target data sets might be distributed only to those who have signed licensing agreements which are legally binding and which prohibit the licensee from: (a) attempting to identify an individual or household in the data files and (b) passing the data to an unlicensed individual.

Snooper access to the target data set could come through several routes:

- Through an authorized data set user.
- In collusion with an authorized user. The collusion could be voluntary (either because the colluder shares the data snooper's goals or has some goal/reward, for example financial payment) or involuntary (if the colluder is threatened or blackmailed).
- Security of a computer containing a copy of the microdata is breached, either by the data snooper or by a hired hacker.
- A copy of the data set is stolen.

In many cases because of legal restrictions, all these access routes involve illegal activity. Therefore, in pursuit of their goals data snoopers must either accept the consequences of their illegal activity; or not release information into the public domain; or release information in a way that conceals the data snooper's identity.

Calculating the probability of such access is prohibitively daunting. However, given the large number of data users, the DSO prudently assumes that the probability of an opportunity is nearly 1. That is, if any individual or organization has decided to attack an anonymized data set, then they will be able to gain access to it. Nevertheless, access to the data for unauthorized usage has potential legal consequences which are likely to lower the probability of an attempt actually being made.

2.1.3.4 Types of Attacks

An attack type is a method for achieving a class of data snooper goals. Five different types of attacks for microdata are identified below plus one for aggregate data.

- *Data set cross-match.* An outside data set with several fields which are identical to or recodable to target microdata record fields (key matching variables) is cross-matched with the target microdata set. The most likely goal for this type of attack would be the enhancement of an outside data set.
- *Match for a single specific individual.* The intention behind this type of attack is to enhance or verify information available about a target individual. Matching information could be from an outside data set—as a hypothetical example, the Inland Revenue in the United Kingdom may want to search the target microdata for income-related information of an individual suspected of tax evasion.
- *Match for single arbitrary individual.* Here the data snooper is not interested in information gathering but instead with being able to claim that identification has been achieved and information could be disclosed. The data snooper is not interested in the actual identity of the individual who has been identified. A journalist in search of a good story might follow this route.
- *Match to a specific group of individuals.* This type of attack is an alternative to Type 3 for certain scenarios and would have the same goal. A set of individuals are selected either because they are distinctive (e.g., they come from a minority ethnic group) or because matching information is available on them (e.g., they belong to an occupation with a register of all members).
- *Fishing.* Strictly, this is not a separate type of attack but a variation on any of the other types. Rather than starting with an individual or a set of individuals in the outside world and attempting to identify them in the target data set, the data snooper starts with the target data set and locates one or more individuals with distinctive characteristics and attempts to find them in the world. Fishing thus starts with a search through the records in a target microdata file. Typically the search seeks to identify records that have unusual combinations of characteristics. The concern is that through demographic knowledge or the analytical properties of the data set, a snooper might be able to identify which records are unique within the population. Having identified such records, the snooper then searches within the population for an individual with those characteristics.
- *Subtraction.* This is mainly a concern for tables of population counts. The principle is that a snooper has knowledge about some individuals within the population which is represented in a given population table. In the simplest case that knowledge comprises all the variables represented in the table. The snooper can then remove the individuals from the table, possibly leading to a table from which disclosive attributions can be made.

2.1.3.5 Key Variables

For all disclosure attempts, key variables are essential to identification. Key variables are those which are available to the data snooper and which are also in the target data set and can therefore allow individuals to be matched. Ideally, the coding of a key variable is identical on both the attack and target data sets (or a harmonized coding can be produced).

There are two sources for key variables: (i) formal data sets containing the same information for the same population and (ii) informal information obtained from local knowledge (e.g., house details obtained via a real estate agent) or personal knowledge (e.g., one's neighbors or fellow workers).

2.1.3.6 Target Variables

The data snooper wants the values of target variables. If the data snooper seeks to gain information, the contents of the target variables are directly relevant. In situations where the data snooper is motivated by the secondary consequences of an attack, identification may be sufficient and therefore the information content of target variables is of little importance. However, in many situations the perceived sensitivity of the information contained in the variable is crucial to the impact of the disclosure on the DSO. Given this, target variables have two central properties, *usefulness* and *sensitivity.*

> *Usefulness*—for a variable to be a target it must contain information which improves upon or verifies data already available to the data snooper.
>
> *Sensitivity*—the sensitivity of a target variable is governed by the perceived importance of the information disclosed for the individual concerned. One might generalize sensitivity as the prevailing perception in the population of the sensitivity of the disclosed information. However, an individual's context will frame the sensitivity of any piece of information. For example, the number of cars I own might be regarded as of low general sensitivity, but an individual who is trying to keep his or her wealth hidden, might regard this information to be of high sensitivity.[2]

2.1.3.7 Effect of Data Divergence

All data sets contain errors and inaccuracies. Respondents do not always supply correct data. There may also be errors in recording. Interviewers make mistakes in recording.

Coders transcribe incorrectly. Oftentimes data items are missing. Missing or inconsistent values may be imputed using methods with no guarantee of accuracy.

[2] There are several likely inputs to both personal and public sensitivity, such as: *perceived deviance* (for example, knowledge that someone does not have an inside toilet could be considered more sensitive than knowledge that they do, because being told of a household that it has an inside toilet does not yield much information against expectations), *social acceptability*, and *information dispersal* (how widely known such information is about the specific individual and about a hypothetical average individual).

Data available from censuses and surveys will be months and possibly years old by the time they are available for analysis outside the DSO. This means that individual and household characteristics will have changed since the date of data collection. An equivalent set of issues will hold for information held by a snooper. The combination of these will create errors in any linkage.

Collectively, we refer to these sources of "noise" in the data as *data divergence*. The term refers to two situation types (i) *data–data divergence*—differences between data sets and (ii) *data–world divergence*—differences between data sets and the world. In general both types can be assumed to reduce the success rate of matching attempts. However, where two data sets diverge from the world in the same way, referred to as *parallel divergence*, then the probability of matching is unaffected. This would be the case, for example, if a respondent has lied consistently or when two data sets both have out-of-date data, but yet have identical values.

2.1.3.8 Likelihood of Success

Likelihood of success is *not* the same as the likelihood of achieving identification given an attack. Rather, it refers to the likelihood of the data snooper achieving their goal, which in some scenarios may not be identification per se (for example, a hypothetical journalist could get a "good story" without a fully verified match).

Goal Achievable by Other Means?

Can the data snooper achieve their goals by other means, which are easier to execute, legal, and/or have an equal or better likelihood of success? This is a crucial factor in determining the likelihood of an attempt being made.

Consequences of Attempt

Each scenario must also consider the likely consequences of an attempt. These consequences will be dependent on the goals of the data snooper and the success or failure of the attempt. Broadly, consequences can be divided into two groups (i) whether or not confidentiality is broken and (ii) the effect on public confidence.

(i) Whether or not confidentiality has been broken

Confidentiality is broken if at least one verified match has taken place, thereby identifying the record in the target data set. A snooper attempt may not break confidentiality, either because a match cannot be achieved with the specified set of variables, or a match cannot be verified. Alternatively, the data snooper may not intend to break confidentiality *per se* but rather to demonstrate that it could be done (as when the individual who is identified has colluded in the matching exercise).

(iia) Knowledge of the attempt to breach confidentiality is released into the public domain.

Whether or not an attempt is successful, information might be released into the public domain regarding the attempt. Release of such information in itself could be dangerous to the DSO because of the effect on public confidence. If the attempt is known to be successful then this impact would, almost certainly, be increased. If the attempt is unsuccessful or demonstrably unverified then the effect is potentially double-edged. The mere knowledge that an attempt has been made might, by bringing the issue to the attention of the public, have a damaging effect on the public's perception of disclosure risk. Also the fact that an attempt has been made indicates that someone believes a breach in confidentiality is possible, which may have an adverse effect on perceived risk. Against this, a DSO could present an unsuccessful attempt as an indication of the security of the confidentiality guarantee. This would depend on the public relations or "fire-fighting" policy of the DSO; see Mackey (2009) for an in-depth discussion of this.

(iib) Knowledge that a breach of confidentiality is possible is released into the public domain.

Here disclosure is demonstrated without a breach in confidentiality taking place (i.e., where the identified individual(s) have colluded in the disclosure exercise). Again the effect on public confidence would depend on the public relations or "fire-fighting" policy of the DSO.

(iic) Details of matched individuals have been released into the public domain.

This is the most damaging consequence in terms of public confidence and future co-operation with the DSO. Where details of matched individuals pass into the public domain, media may run a "personal story" which may magnify (in the public mind) the importance of the confidentiality break. This will be so even if the match is unverified.

Additional damage may be done if the information disclosed is sensitive or personally embarrassing to the matched individual.

Likelihood of Attempt

Given the key variables, likelihood of attempt is a critical output of the scenario analysis. Surely it is difficult in general to put a numerical probability to the likelihood of a data snooper attack. Many of the inputs into our classification are highly contextual. Also, in order to make any assessment at all we must assume some rationality on the part of the would-be snooper. However, it is usually possible to arrive at a conclusion/exclusion decision by making this assumption and then to review the input and process information in each possible scenario.

Effect of Variations in Data Set Structure

Variations in the target data set, for example in content, sample size, or geographical detail, may alter the likelihood or meaning of a particular scenario.

2.1.4 Data Environment Analysis

Scenario analysis as described in Section 2.1.3 is important in understanding disclosure risk. However, the Achilles' heel for DSOs in dealing with the practicalities of risk assessment is their lack of knowledge about what data snoopers know. Do they have access to data that are external to an intended release that could be used in compromising confidentiality? This in particular makes it difficult for the DSO to generate an informed list of key variables. Purdam et al. have developed a method called *data environment analysis* to help rectify this (Purdam and Elliot, 2002; Purdam et al., 2003b, a). In data environment analysis, forms used to collect information about individuals are decoded into metadata. Corresponding to that form, a record is created on the data environment metadata set which records all the information obtained from the form plus, where possible, an estimate of coverage. Coverage information is obtained from common sense estimates of service value, interviews, questionnaires, and public statements of data set holders. This data environment analysis enables input into scenario generation. Elliot et al. (2005) have outlined a design using grid technology to automate Data Environment Analysis as part of a larger system of automated statistical disclosure risk analysis.

2.2 Assessing the Risk

Having outlined the concepts underlying data snooping, we now consider more fully the practice of risk assessment, which is a core concern of a DSO. Much early research was data-centric, that is it focused on defining properties of a data release that were more or less risky. This leads to the concept of uniqueness.

2.2.1 Uniqueness

Underpinning much of the work on disclosure risk analysis, particularly for microdata, is the notion of *uniqueness*. A record is unique on a set of key variables if no other record shares its values for those variables. We need to examine two types of uniqueness: *population uniqueness*—a unit is unique in the population (or within a population data file such as a census) and sample *uniqueness*—a sample unit is unique within the sample file. These two concepts are the basis for many of the disclosure risk assessment methods for microdata. If a unit is population unique then disclosure will occur if a snooper knows it is unique. Sample uniqueness is a necessary precondition of population uniqueness. Much existing methodology concerns using sample information to make inferences about population uniqueness. These methods are described in more detail in Chapter 3.

2.2.2 Matching/Reidentification Experiments

Using the same methods that a data snooper would, matching or reidentification studies simulate the linking of records from an identification file with those on the target microdata file (Elliot and Dale, 1998; Müller et al., 1995; Winkler, 1995a; Elliot, 2007). Such studies have the advantage that they are generated by empirical data, rather than depending on theoretical values provided by the uniqueness statistics discussed in the last section. However, there are corresponding disadvantages:

(i) We cannot be certain that a particular identification data set will provide a generalizable measure of the level of disclosure risk associated with the target microdata set; the results will be specific to the identification data set and the particular experiments conducted. A different data set with different data divergence from the target microdata set might well produce substantially different results.

(ii) Setting up matching experiments is time consuming, with considerable effort usually required to arrive at a harmonized coding for the two data sets.[3] Complicated procedures are usually necessary to verify accuracy.

2.2.3 Disclosure Risk Assessment for Aggregate Data

Much of the work on disclosure risk assessment has focused on modeling identification risk in microdata. Little has been done on exploring and developing equivalent risk assessment metrics for aggregate data. One reason for this imbalance is that whilst the conceptual structure of attacks on microdata (identification-attribution) is established and basically understood the equivalent conceptual structure for aggregate data (subtraction-attribution) is not.

Generally, risk assessment for tabular data has used ad hoc proxy measures. For frequency data, one commonly used measure is based on the numbers of "small" cell counts. For magnitude tables, a measure which identifies cells with dominant respondents leads to the heuristic *p*/*q* rule. Smith and Elliot (2008) provide a more theoretically grounded algorithm known as the Subtraction-attribution-probability (SAP), which generates a probability of a snooper being able to recover one or

[3] Although these are the same processes data snoopers would have to go through if they are to attempt a confidentiality breach.

more zeroes in a table, given specified knowledge about the population. One of the advantages of this metric is that it can be applied equally to unperturbed or perturbed tables, to single or sets of tables, and even to unreleased cross-classifications of released margins (effectively dealing with the linked tables problem). The method is described in more detail in Chapter 3.

2.3 Controlling the Risk

Having appropriate methods for measuring disclosure risk, the DSO must consider what to do about the measured level of risk. If the assessed risk is too high, it has two options: restrict access to the data or restrict the data. Often both of these options are pursued. In this section we consider several ways of controlling disclosure risk.

2.3.1 Metadata Level Controls

Controls at the metadata level work with the overall structure of the data release. The key components of such controls are the sampling fraction, choice of variables, and level of detail on those variables.

Sampling Fraction

For surveys, the sampling fraction is specified by the study design and so its choice often rests outside disclosure control. Nevertheless, the sampling fraction is critical in determining disclosure risk for a microdata file.

Choice of Variables

An obvious mechanism of disclosure control is by excluding certain variables from the released data set. The DSO can (i) reduce the number of key variables—those which a plausible snooper is likely to have access to or (ii) reduce the number of target variables. These choices flow naturally from the scenario analyses described in Section 2.1.3. With microdata, the choice is whether a variable appears in a data set or not. With aggregate data, the choices are about which variables will be included in each table.

Level of Detail

Decisions over level of detail mirror those over choice of variables. Here the DSO will look at categories with small counts and determine whether merging them with other categories would significantly lower disclosure risk without losing appreciable information. Not surprisingly, many data users would like the maximum level of detail possible on every data set. But DSOs regard some variables, especially geography and time, as particularly problematic in maintaining confidentiality. Area of residence is a highly visible component of an individual's identity, therefore geographical detail is often constrained and data are released at coarser detail than data users would like. Similarly, sometime variables, such as exact date of birth, can when combined with other variables be straightforwardly identifying.

2.3.2 Distorting the Data

The main alternative to metadata controls is various forms of data distortion, which we call *perturbation*. These techniques manipulate the data in order to foil any identification/subtraction strategy so that a snooper cannot be certain of any match or recovered zero. In this section we will look at several methods of perturbation that are commonly used for disclosure limitation.

Data swapping involves moving data between records in a microdata set. A particular form of this, often called "record swapping," involves swapping the geographical codes of two records. Data swapping will be discussed in detail in Chapter 5.

Rounding is a technique used with tables of counts. In the simplest form all the counts are rounded to the nearest multiple of a base (often three, five, or ten). Counts which are a multiple of the base number remain unchanged. Normally, the margins are rounded according to the same method of the internal cells. Therefore, in many cases this method does not yield an additive table. This fact has motivated rounding variants that are described in Chapter 4.

For magnitude tables, *controlled tabular adjustment* or *cell perturbation* have been suggested. These techniques modify the original values in the table, but typically only the internal cells. In this way, marginal values are precise whilst the internal cell values are uncertain. Again these techniques are described in Chapter 4.

Cell suppression is an SDL technique that can be implemented in various forms whereby the data are only partially released. In one sense, releases of aggregate data are themselves primary examples of suppression, since they are partial releases of the underlying microdata (or full table). If I release two one-way marginal frequency tables, but not the joint table, I am suppressing the cross-classification. In the cell suppression approach which is described in Chapter 4 individual cells are suppressed according to specified rules. Suppression can also be used in microdata where particular variables can be suppressed for particular cases.

2.3.3 Controlling Access

The DSO can control who may access the data and the manner of that access. In practice, only coarse aggregates are released without any restriction. With microdata, licensing is typically required to get access. Control over who accesses the data, for what purposes, and by which medium is often used in combination with SDL techniques. For example, the Australian Bureau of Statistics (ABS) has released microdata from its census at three levels of detail. The least detailed is released on CD to users under license, a more detailed version is accessible via the Internet through ABS's remote access data laboratory, and the most detailed version is only accessible to trusted users who must do their work at ABS's offices.

2.4 Data Utility Impact

The final piece of the jigsaw in building an understanding of the risk assessment domain is the notion of *data utility*. The concern is that disclosure limitation will not only stymie a would-be data snooper, but also make them unusable for intended users.

The common approach to data utility is to generate metrics for the *information loss* caused by the disclosure control employed with a given data set. This can be either based on maintaining certain key statistics, for example, means, variances, co-variances, and so on (Cox et al., 2004; Domingo-Ferrer and Mateo-Sanz, 2001) or by maintaining some construct such as Shannon entropy or uncertainty functions (Duncan and Lambert, 1986).

The advantage of these approaches is that they are easily replicable and comparisons can be made between data sets. The major drawback of the information loss approach to data utility impact is that it is difficult to relate it to the actual utility of the data because data users are ultimately interested only in the data's usability for the analyses that they intend to conduct. If the data are fine for that purpose then the user is indifferent to what the DSO has done to them. Conversely, if they are not fit for the data users' purposes, then the DSO has contributed nothing of value, even if according to some measure there has been no substantial information loss.

The impact on data utility of SDL techniques falls into two categories, *reduction of analytical completeness* and *loss of analytical validity* (Purdam and Elliot, 2007). With some SDL methods, typically metadata controls, analyses that could have been conducted cannot be. This is a reduction in analytical completeness. Use, for example, of geographical thresholds in microdata sets leads to smaller administrative units being grouped together, thereby preventing researchers concerned with inferences about the smaller units from using the data set effectively. The loss of analytical validity is harder to define, but in some ways more critical because of its insidious nature. Loss of validity occurs when an SDL method has altered a data set to the point where a user reaches different conclusions from the same analysis. This is a danger with perturbative SDL techniques.

Removing a single variable from a data set may have an impact on the quality of the data as measured in an information-theoretic sense, but might have no effect on analytical completeness, certainly if no user would use that variable. Conversely, a minor perturbation in information-theoretic terms could have a significant effect on analytical validity if the perturbation affects important variables disproportionately.

Given the problems of the information-theoretic approach, Purdam and Elliot (2007) have developed retrospective methods for more directly assessing the impact on data utility. They survey data users (typically authors of studies) to assess the impact on analytical completeness. They also replicate published studies after the application of disclosure control techniques to assess the impact on analytical validity. These allow a more direct analysis of utility but suffer from the same problem of non-generalizability as the direct record linkage studies of risk.

In Chapter 6 we examine data utility in more detail describing notions such as the Risk-Utility Confidentiality Map.

2.5 Summary

As we have described it, the SDL field is comprised of three areas: (i) disclosure risk analysis (the assessment of the risk), (ii) disclosure limitation techniques (to reduce the risk); and (iii) the analysis of the impact on the utility of the data of those limitation techniques. Each of these areas poses interesting and challenging academic questions. However, development of SDL methodology is not solely or even primarily academically driven. One of the features of the SDL field is that as the academic questions are addressed they lead to changes in procedures within DSOs, usually first in the national statistical institutes. Indeed, many of the key researchers in the field are employed by the National Statistical Offices rather than by academic institutes. This combination of theoretical interest with policy and practice relevance makes SDL a compelling field for both researchers and practitioners of statistical confidentiality. Demonstrating this richness in both theory and in practice, the next chapter delves more deeply into the issues of disclosure risk analysis.

Chapter 3
Assessment of Disclosure Risk

Before disseminating a data product for public use, a DSO needs to assess the risk of a data snooper compromising confidentiality. In its original form as the source data, a data product typically has unacceptably high disclosure risk. The data product must therefore be transformed to lower the disclosure risk to an acceptable level. We present a variety of methods for statistical disclosure limitation in Chapters 4 and 5, but first we need to understand disclosure risk and have appropriate tools for its assessment.

To begin the assessment of disclosure risk in a way that supports rational decision making, the DSO should be cognizant of three components: the various possible outcomes of data snooper actions, the (dis)utility of those outcomes, and the likelihood of those outcomes occurring.

We cannot expect a DSO to have perfect knowledge of any of these components because of three complications:

1. Disclosure risk is dependent on the data environment—the context of the data release. This includes the purpose for which the data set was collected, the methods of data collection, how the data are stored, the details of planned data release, and the motivations, capabilities, knowledge, and resources of potential data snoopers. The impact on disclosure risk of the data environment is difficult to assess and constantly changing.
2. A data snooper will seek to link the identification information they hold to the anonymized data. But typically there is a difference between the information stored in a data set about a given population unit and that which is available to a data snooper. This difference is called *data divergence*, and it has a substantial impact on the ability of a data snooper to link records. Necessarily the DSO is uncertain about the extent of data divergence. Furthermore, the DSO cannot be sure about how much variation in data divergence there is between different pairs of data sets. Also, just how is it distributed within a data set? Are, for example, higher risk records subject to more (or less) divergence?
3. Estimates of the (negative) utility of a disclosure are typically based on little evidence. We do not know how severe the impacts of a disclosure would be. Furthermore, utility is itself affected by a complex, dynamic system which

G.T. Duncan et al., *Statistical Confidentiality*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-7802-8_3,

includes the reactions of the DSO itself, other organizations, the media, and public opinion. Even the implicit belief that the utility associated with a disclosure event must be negative is an assumption, not an empirical fact. Indeed, it is possible that a disclosure, if exceptionally well-handled by the DSO, might actually lead to net positive utility.

In the face of these complications what can a DSO do? The standard approach is to assume that the disutility associated with even a single disclosure is high, but also to assume only a moderate level of knowledge on the part of the data snooper.

Many DSOs now engage in scenario analysis—like the method developed by Elliot and Dale (1999)—which provides them with a conceptual model of where an attack might come from and what resources a data snooper might employ. Crucially in this, the DSO specifies the *key variables*—the variables that data snoopers could use to carry out their attack.

This chapter starts from where the DSO has determined the form of attack that it wishes to protect against. The DSO wants to obtain some measure of the likelihood of such an attack being successful given that it has occurred. Scenario analysis sorts "unlikely" from "likely" attacks. Consistent with our discussion in Chapter 2, we employ Marsh et al.'s (1991) equation:

$$P(\text{identification}) = P(\text{identification}|\text{attempt}) \cdot P(\text{attempt}).$$

We treat P(attempt) as a binary variable with the likely scenarios having a value of 1 and the unlikely ones a value of 0. While we will discuss this simplification further at the end of the chapter, for now, we use it and examine various ways of assessing risk.

3.1 Thresholds and Other Proxies

Early on, disclosure risk measures often used population thresholds. The risk would be deemed to be unacceptably high (some statistic fails a threshold test) or acceptably low (it does not). For example, the minimum expected population counts for the 1991 UK census were set for each univariate category at 25,000 for individual microdata, based on the formula

$$C = (1/X) \times (Y/Z),$$

where C is an expected count, X is the sampling fraction, Y is the national population, and Z is the geographical threshold. See Dale and Marsh (1993).

Depending on the type of disclosure risk measure that is used and the level of protection that is needed in the publication, a threshold, say τ, is set. If the risk is below τ, the data product can be released; if the risk is above τ, more disclosure limitation is necessary. The level of acceptable risk depends on how the data product is to be disseminated. Data released for on-site data laboratories may need different protection levels from that in licensed data archives which in turn may differ from that required for public use. Such threshold rules are simple to understand and to implement. The disadvantage with them is that they are not directly related to what a data snooper may do. They are akin to estimating the risk of a burglary at a warehouse based on the value of the items contained there but without considering the security employed. Since such crude proxies are inadequate, more sophisticated measures are needed.

We now discuss disclosure risk measures, beginning with those derived from matching microdata files.

3.2 Risk Assessment for Microdata: Types of Matching

As Chapter 2 describes, microdata are at risk because a data snooper may use some sort of matching process of records in the subject file and an identified file available to the snooper. Directly or indirectly, all metrics for microdata risk measure the extent to which matching is possible. Should we measure risk for each individual record, or overall for a data file? Certainly we can measure the extent of risk either at the file level or at the record level. We next explore how each can be done and what the relative merits are.

3.2.1 File-Level Risk Metrics

File-level risk metrics measure the average risk across the whole data file. In some cases, they give a useful overview of the data file risk. DSOs find this worthwhile in determining whether recodes of key variables are beneficial. Many file-level metrics are based on the concept of population uniqueness.

3.2.1.1 Population Uniqueness

As noted in Chapter 2, uniqueness is an important aspect in understanding disclosure risk for microdata. A type of uniqueness, *population uniqueness,* was considered early in the discussions of disclosure risk. See Dalenius (1986) and Greenberg (1990). If an individual has unique values on a set of key variables within a population, then that individual is said to be a *population unique*. The proportion of such individuals in a given population is the level of *population uniqueness*.

Population uniqueness has advantages of simplicity and an intuitive relationship with disclosure risk. If an individual is known to be a population unique and a record matching that individual is found within a data set, then there is high probability that identification disclosure can occur. However, population uniqueness has several disadvantages as an underlying concept for disclosure risk:

1. Determining population uniqueness requires access to population data. This is rarely available outside National Statistical Institutes—except in a limited and

incomplete form (in some cases electoral registers may approximate population coverage). Certainly, for data collected through sample surveys, the DSO will not have access to directly correspondent population data.

2. Population uniqueness has no logical connection to the fraction of the population surveyed, and yet the sampling fraction is an established factor of disclosure risk for an individual data unit.
3. There is no obvious method of incorporating *data divergence* in the estimate of risk. Clearly, data divergence caused by coding differences, response errors, measurement errors and so forth does affect risk levels, but there is no principled method for adjusting a population uniqueness statistic to incorporate the effect of divergence.

Bearing in mind these disadvantages, many researchers have found the advantages compelling enough to try to develop methods for overcoming them. Most of this work has addressed point 1 above, about the DSO not having population data. They have tried to infer something about population uniqueness from the information they do have, namely that a data unit is unique in the sample. In the process, they also help to overcome point 2 above about the sampling fraction.

3.2.1.2 The Proportion of Sample Uniques that are Population Unique

What is the probability of a population unit being a population unique on a set of key variables given that it is unique in a sample on those same variables? This conditional probability is now accepted as a more useful measure of disclosure risk than the level of population uniqueness because it is sensitive to changes in sampling fraction in a monotonic way. Nonetheless, it shares problems with population uniqueness:

1. Access to population data is required to calculate it.
2. Although the conditional probability is sensitive to sampling fraction, it does not explicitly incorporate the sampling fraction within its calculation. Clearly, there is a change in absolute risk that is directly proportional to the probability of an individual being in the sample data set and this is not part of the calculation of this statistic.
3. Its relationship with geographical scale can be non-monotonic. Intuitively, as geographical scale decreases, disclosure risk ought to increase. However, Elliot et al. (1998) found a surprising and unusual result that for certain UK data the conditional probability had a non-monotonic relationship with population size for geographical units of between 30,000 and 1,000,000.

3.2.1.3 The Skinner and Elliot Method

An alternative approach developed by Skinner and Elliot (2002) overcomes some of the above problems. This approach focuses on the probability of a correct match given a unique match. For microdata, Elliot (2000) observed that a data snooper attack could be mimicked by the following three-stage process:

- Remove a record from a data set.
- Conditionally replace it with a probability equal to the original sampling faction.
- Match the removed record against the data set (on a selected set of key variables X).

Furthermore, we can generalize this process by observing that there are six possible outcomes as shown in Table 3.1. The two bold cells in this table are critical. One is where the record was a sample unique on X and copied back into the file yields a correct unique match. The other is where the record was one of two with the same values for X and was not copied back yields a false unique match.

Table 3.1 Possible outcome of removing a record from a microdata sample and then copying back at a probability equal to the original sampling faction and finally matching the removed record against the file; after Elliot (2000)

Record is:	Copied back	Not copied back
Sample unique	**Correct unique match**	Non-match
One of a sample pair	Multiple match including correct	**False unique match**
One of a larger equivalence class	Multiple match including correct	False multiple match

Therefore, we can obtain an estimated probability of a correct match (cm) given a unique match (um) through the sums of the number of records with sample frequencies of 1 and 2 and the sampling fraction π:

$$P(cm|um) \cong \frac{\sum_j I(f_j = 1)\pi}{\sum_j I(f_j = 1) + \sum_j I(f_j = 2)(1 - \pi)}.$$

In the above expression, f_j is the frequency of the values on K possessed by record j and $I(x)$ is the indicator function which is 1 when x is true and 0 when it is false.

Skinner and Elliot (2002) further develop this strategy for various sample designs and report a numerical study showing that the method produces accurate estimates of the probabilities. They also obtain an estimator for the standard error of the predicted probability.

Although useful, because it does not require population data and mimics more closely what a data snooper might do, the Skinner and Elliot solution has two problems:

- It does not take account of data divergence between the data snooper's file and the target file. Therefore, it always overestimates the probability.
- The probability it generates is based upon a scenario where a data snooper draws a unit randomly (so with equal probabilities) from the population and

then matches those against the target file. This scenario does not take account of information in the target file which might focus the data snooper on higher risk records.

It is now generally accepted that risk analysis at the file level, whilst useful in such things as determining levels of coding, and so on, only provides a partial measure of identification risk for cross-match attacks. To assess disclosure risk for other attack methods we need a record-level measure of risk.

3.2.2 Record-Level Risk Metrics

3.2.2.1 Probability Modeling Approaches

Intuitively, records presenting combinations of key variables that are unusual or rare in the population have high disclosure risk, but rare or even unique combinations in the sample do not necessarily correspond to high risk individuals. To develop this intuition further, consider a cross-classification of key variables. Each cell in the cross-classification is the cross-product of the categories of the key variables. Let F_k be the number of individuals in the population which belong to cell k, and let f_k be the given sample frequency of this cell. We define $1/F_k$ as the probability of reidentification of an individual in cell k. We then need to infer the population frequency F_k from the sample frequency f_k. There are two main methods for risk assessment. One is the *Poisson Model*, developed by Skinner and Holmes (1998) and refined by Elamir and Skinner (2006). It is based on the assumption that $F_k|f_k$ has a Poisson distribution. The other is the *Argus Model* developed by Benedetti et al. (2003), and further developed by Polettini and Stander (2004). It is based on the assumption that $F_k|f_k$ has a Negative Binomial distribution. In both methods, the individual risk measures can be aggregated to obtain global risk measures for the entire file. The risk is defined as the global measure $E[1/F_k|f_k]$.

Skinner and Holmes (1998) consider a data snooper who attempts to disclose information about a set of identified population units, which they term "targets." The snooper is assumed to have prior information about the key values of the targets and attempts to establish a link between these and individual records in the released microdata file using the values of the key attributes. Skinner and Holmes assume that the snooper finds that a record r in the microdata file matches a target with respect to the key X. Now F_i is the number of units in the population with $X = i$ and we let $i(r)$ denote the value of X for record r. If $F_{i(r)}$ was known, the snooper could infer that the probability of a correct link is $1/F_{i(r)}$ and if $F_{i(r)} = 1$ the snooper could infer the link is correct with certainty.[1] However, usually the snooper will not know the true value of $F_{i(r)}$ since the microdata set contains only a sample, but by introducing a super-population model the snooper could attach a probability distribution $P(F_i = j)$

[1] Assuming no data divergence.

to the cell frequencies. Extensions of this approach are Skinner and Shlomo (2008) and Elamir and Skinner (2006).

3.2.2.2 Special Uniqueness

An alternative approach developed by Elliot et al. (2002) uses heuristic measures based on the concept of a *special unique*. Consider a data set that is a sample of population units cross-classified on at least two variables. Employ the following notation:

X is a set of key variables
Y is an arbitrary set of variables such that $Y \subset X$
$f_i =$ number of data units with $X = i$
$f_i =$ number of data units with $Y = f_j$

A data unit is called special unique with respect to X if $f_i = f_j = 1$.

Elliot et al. (1998) observe that a special unique has a higher probability of being a population unique than a sample unique which is not special (which they refer to as a *random unique*). The unique subset components of a special unique are called *minimal sample uniques* (MSUs), that is, one which contains no unique subsets.[2,3]

Elliot et al. (2002) show that as the size of the MSUs decreases and the number of the MSUs increases for a given data unit, so the probability of that data unit being population unique increases. Thus, if it is possible to obtain a metric which collates for a given data unit and set key variables the size and number of MSUs, then we may have a way to calculate the risk for each data unit. Implementing this observation in a principled way is more difficult. The method that Elliot et al. (2002) implemented into SUDA (Special Uniques Detection Algorithm), which they term the “SUDA score,” is based on the principle that the key variable values form a lattice structure as shown in Fig. 3.1.

The SUDA score exploits the fact that the smaller an MSU, the larger the number of paths through the lattice that are unique, and therefore the larger the proportion of the lattice that is unique. In essence, the SUDA score is the sum of the number of paths through the lattice that are unique. Experiments show respectable correlations between this score and the underlying risk measure $1/F_k$.

Elliot et al. (2002) combined the SUDA score with the Skinner and Elliot measure described in Section 3.2.1 to obtain a pseudo-probabilistic measure of data snooper confidence in a given match.

[2] In fact, Elliot et al. (2002) reverse the sub/superset terminology as they are looking at it from the perspective of a search through a lattice starting with the null set.

[3] It is worth noting that each special unique might contain any number of MSUs and each record may contain multiple special unqiues – i.e. different combinations of variables which are special unqiue. It also worth noting that, by definition, MSUs are random uniques.

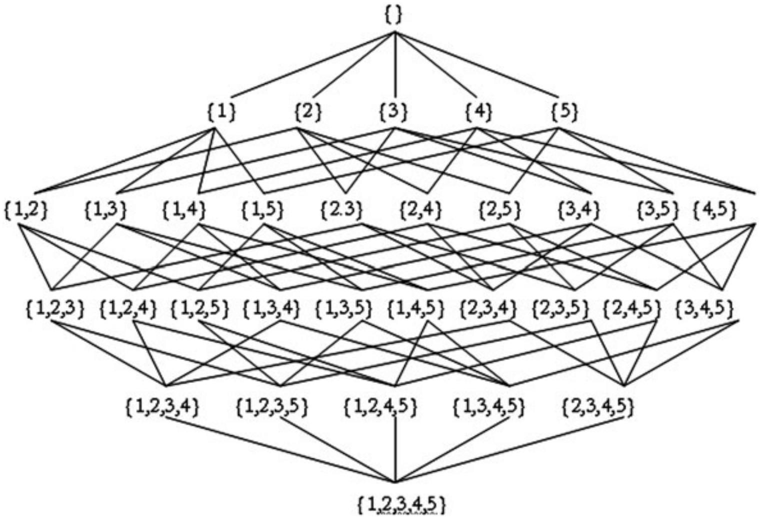


Fig. 3.1 Lattice for $\Xi = \{1, 2, 3, 4, 5\}$; from Elliot et al. (2002)

Again these approaches are heuristic, and like many such techniques it is impossible to specify the conditions under which they break down. However, Elliot et al. (2002) were able to demonstrate sufficient robustness that the technique was used with 2001 UK census microdata.

3.3 Record Linkage Studies

Most of the measures discussed in this chapter so far assume data divergence is zero. That is, they are calculated as if the prior information that the data snooper has about a given population unit will be the same as the information the data snooper has about that population unit in the target data set. Because this assumption is false in general, it leads to overestimation of the disclosure risk. Although this can be accounted for in an *ad hoc* manner, this overestimation becomes a problem when we want to assess the impact on disclosure risk of deliberately introduced divergence such as perturbation occasioned by SDL methods.

As described earlier, the risk of identification disclosure is taken to be the probability that a data snooper can correctly identify a record in the released data set. Most of the risk measures discussed above are based on the principle that the snooper seeks to link using identification information as in Fig. 2.1. We can get an estimate of this probability by attempting to link a record in a second data set with a record in the data to be released. Furthermore, we can attempt this linkage before and after the SDL to assess the impact of the SDL on the data set to be released. This approach will enable us to estimate the impact of the SDL.

We explore two ways to implement this approach. The first is to make use of an available data set (either the one that the DSO owns or the one that is publicly available). Plausibly, such data sets include those used in direct marketing as well as government records such as driver's licenses and voter registration. This is an approach taken by Paass (1988), the Federal Committee on Statistical Methodology (1994), and Yancey et al. (2002). The matching is attempted through record linkage software which looks for similar records in the external data set and the target data set. The second approach is to use the original data set, matching records in this unperturbed data set with records in the masked data set. Since the data snooper does not have access to the original data set, a DSO needs to use scenario analysis as described in Section 2.1 to generate various assumptions about the knowledge of the data snooper and thereby select the key variables.

3.3.1 Using an External Data Set

In an early study, Müller et al. (1995) matched the 1987 North Rhine-Westphalia microcensus with a handbook of German scientists and academics. Using ten key variables and simple matching techniques, only four records (representing 5% of the entries in the handbook) were correctly matched.

Elliot and Dale (1998) report on linkage of the 1991 UK census microdata with microdata from the General Household Survey (GHS). These two data sets were harmonized to generate a key variable set of 19 variables. Only data for the GHS which were collected during the month of the census were used. The matches were verified by ONS staff. In this study there were 6 correct unique matches and 219 incorrect unique matches; giving a correct match rate of only 2.7%. By using an early version of the Skinner and Elliot method, Elliot (2000) was able to calculate the theoretical match rate to be 8.8% for these two files. This is interesting because it indicates a high level of data divergence between the two files despite the efforts at harmonization.

In a similar study Elliot (2007) tested the impact on disclosure risk of the SDL carried out on the 2001 UK census microdata (the SARs). Matching was attempted against the pre- and post- SDL SARs using the UK Labour Force Survey (LFS). Again effort was put into harmonizing the two data sets, across twelve key variables. This time however a more sophisticated snooping attack was simulated with the attack taking into account the higher risk records as identified by Elliot et al.'s (2002) Special Uniques system. Overall correct matching rates were similar to the 1998 study, with the match rate between the LFS and the pre-SDL file being about 2.7% and the post-SDL file of about 2.2%. However, these results did not reveal either the variation in risk or the impact of the SDL. By focusing only on those records that the special uniques program indicated were high risk, a correct matching rate of over 20% was achieved with the pre-SDL file. This was reduced to 6–8%

on the post-SDL file (depending on what level of the special uniques metric was used as a threshold). This indicated both the importance of the variation of risk across files and also how the use of targeted SDL impacts on that risk.

3.3.2 Using the Pre-SDL Data Set

A criticism of using external data sets is that results are dependent on the choice of data set and therefore generalizability is precarious. Even in comparison of pre- and post-SDL data sets, which may appear to give results which are independent of the external data set, we cannot be sure that the choice does not impact on the results because natural data divergence probably interacts with that provided by the SDL. To get around this difficulty, some researchers have taken the external data set out of assessing the impact of disclosure by linking back to the pre-SDC data set itself. The result is more generalizable as we are no longer at the mercy of interaction between data divergence and the impact of SDL techniques.

Elliot (2001) provides a variant to the file-level method developed by Skinner and Elliot (2002) which allows for known false matches. This allows the assessment of the impact of SDL that has been applied to the data on the file-level probability that Skinner and Elliot's method produces.

3.3.2.1 Distance-Based Record Linkage

When the whole of the data set is released in some perturbed form, disclosure risk is a function of the extent of perturbation. Just how depends on the mode of snooper attack. Consistent with Domingo-Ferrer and Torra's (2003) argument that record linkage is the most general and realistic attack mode, in this and the next section we assume attack by record linkage (see, e.g., Winkler, 1995 a, b).

As in any record linkage context, the data snooper has an external data set containing identifiers and the same attributes that are present in the masked and the released data set (Gill, 2001). The snooper tries to match an identified record from the external data set with a record in the perturbed confidential data set. Once some records are linked in the masked data set with records in the external data set, confidentiality has been breached because an identity disclosure has taken place (see, e.g., Bacher et al., 2002).

Typically the DSO will not have access to, or even know about, all external data sets that a data snooper might employ. Therefore a conservative way—perhaps far too conservative way—to assess disclosure risk is to see how well the original data can be linked to the masked data. Further, as noted by Winkler (1995a), true matches in the original and external data may differ on key variables because of errors in the records. The original data then are a surrogate for an *ideal* (from the snooper's viewpoint!) external data set. Disclosure risk is measured by the extent to which the DSO can reidentify records in the masked data. Since perturbation of key variables can upset the snooper's linkage agenda, the extent to which this happens is inversely related to disclosure risk. This approach to disclosure risk assessment is followed by Pagliuca and Seri (1999) using distance-based record linkage.

In distance-based record linkage, the first step is computing the "distances" between records in the original data set and records in the masked data set. This requires some appropriate metric, such as for continuous variables, Mahalanobis distance (Torra et al., 2006). For those with a more mathematical bent who are interested in the mechanics, the following material provides the details of this Mahalanobis distance implementation.

Let the *i*th original record be $\mathbf{X}_\mathrm{i}$ and the masked record be $\mathbf{X}_\mathrm{i}^\mathrm{M}$. The Mahalanobis distance between the original record and the masked record is

$$\Delta(\mathbf{X}_\mathrm{i}, \mathbf{X}_\mathrm{i}^\mathrm{M}) = (\mathbf{X}_\mathrm{i} - \mathbf{X}_\mathrm{i}^\mathrm{M})^T[\mathbf{S}_O + \mathbf{S}_M - 2\mathbf{S}_{OM}]^{-1}(\mathbf{X}_\mathrm{i} - \mathbf{X}_\mathrm{i}^\mathrm{M}),$$

where $\mathbf{S}_O$ is the computed variance matrix of the original data set,
$\mathbf{S}_M$ is the computed variance matrix of the masked data set, and
$\mathbf{S}_{OM}$ is the computed covariance matrix of the original and masked data sets.

There is ambiguity in how to compute the covariance matrix $\mathbf{S}_{OM}$, since it is not obvious which records in the original data set should be matched with which records in the masked data set. One resolution of this ambiguity is the following.

For each record in the masked data set:

1. Calculate the distance to every record in the original data set.
2. Identify the "nearest" and "second nearest" records in the original data set.
3. Label the record as "linked" when the nearest record in the original data set has the same record number as the corresponding original record. Label a record in the masked data set as "linked to 2nd nearest" when the second nearest record in the original data set has the same record number. In all other cases, label a record in the masked data set as "not linked."

The percentage of "linked" and percentage of "linked or linked to 2nd nearest" are measures of disclosure risk. Empirical studies suggest that setting the covariance matrix $\mathbf{S}_{OM}$ equal to zero produces a higher number of reidentifications. Also see Domingo-Ferrer et al. (2006). Domingo-Ferrer (2002) explores distance-based record linkage for categorical variables, including both ordinal and nominal variables.

3.3.2.2 Probabilistic Record Linkage

The basic idea of probabilistic record linkage is to use probability models to assign an index value $I(a,b)$ to every pair of records (a,b) where a is in file A and b is in file B. Typically, the index $I(a,b)$ is computed using conditional probabilities as a log-likelihood ratio defined by $I(a,b) = \log\left(\frac{P(a=b)|(a,b)\in M}{P(a=b)|(a,b)\in U}\right)$, where M is the set of true matched pairs and U is the set of true unmatched pairs. Probabilistic record linkage is introduced in Fellegi and Sunter (1969). Jaro (1989) provides an application that illustrates its effectiveness. Winkler (1995a, b) describes further developments of the theory.

Domingo-Ferrer and Torra (2001, 2002) compare probabilistic record linkage to distance-based record linkage. The idea is to obtain a measure of how well the data snooper can reidentify the masked records through linkage of the original data and the masked data. They suggest that the two methods give comparable results, so either can be used as a measure of disclosure risk.

3.4 Risk Assessment for Count Data

Risk assessment for count data is a relatively undeveloped area. The simplest approach is to consider low counts to be problematic and to set some threshold below which unperturbed low counts are deemed to be problematic. Seeking a more theory-based approach, Smith and Elliot (2005, 2008) argue that risk of attribution exists if, and only if, one or more zeros exist in some population cross-classification.[4] Information-theoretically the following can be considered axiomatic: if we know that an individual is in a population and that a cross-classification of that population across a given set of variables contains a zero, then we know that that population unit does not possess the combination of features for which the zero is recorded.

In general, a set of population cross-classifications can be used to place bounds on any cross-classification from which they could be derived. It is enough to consider only the "base" cross-classification (i.e., the full table constituted from all of the variables in the released cross-classifications). Any cross-classification over a superset of the variables in the base cross-classification contains (recovered) zeros if, and only if, the base cross-classification contains (recovered) zeros. Bounds on smaller margins can be solved, but again this is unnecessary, as any zero in a margin implies zeros in the full cross-classification.

Smith and Elliot (2008) propose a measure based simply on the presence of zeros in the full population cross-classification. The crucial point of departure with other treatments of disclosure risk for tables of counts is that Smith and Elliot explicitly take into account the data snooper's knowledge in their formulation. Specifically, their measure is the "probability of recovering one or more zeros in the full cross-classification given the subtraction of a random sample of n population units," according to the principle illustrated in Tables 2.2 and 2.3. They term this as the *subtraction attribution probability* (SAP). It is the probability of recovering a zero within a population cross-classification after the removal of the information for any identified data units for which all data values for variables within the cross-classification are known.

[4] Such a population cross-classification need not necessarily have been released. For example, it is possible to construct examples where the exact counts in a 3-way cross-classification can be recovered from its three distinct 2-way margins.

To understand how this works, consider a base table of counts with cell counts c_i, $i = 1$ to m. Now, given a published set of marginal tables, with each cell independently perturbed, each published count, x, will imply a pair of constraints of the form, $l \leq c$, $c \leq u$, where l and u are the trivial bounds (that is the minimum and maximum possible real value given what the observer knows about the perturbation scheme).

In such a situation, the recovery of a zero by subtraction of a known sample of the population occurs if, and only if, the sample (that is the data for known individuals already held by the data snooper) implies that $s_i = c_i = u'_i$, where s_i is the corresponding known sample count and u' is the table of the tightest upper bounds on the base table implied by the set of all linear constraints.

To illustrate, take a simple example. Suppose that a simple rounding scheme rounds all values to base 5, so values 0–2 will be published as "0", 3–7 as "5", 8–12 as "10" and so on. Therefore, a snooper observing a value of 5 in such a table of counts knows that the true value lies between 3 and 7. The snooper may also, by considering other values in the table, be able to tighten the bounds further (i.e., reduce the range of possible values) but let us suppose that s/he cannot. Let us consider further that the intruder happens to have data for seven individuals with that combination of characteristics. Since 7 is the maximum value for that cell s/he knows it is the actual value and can subtract those seven individuals from the table and therefore will have recovered a 0.

In general, the probability of a snooper recovering at least one zero for a given level of knowledge (number of individuals with known values), equivalent to a random sample of size n, can be computed using the formula:

$$SAP(n) = \frac{\sum_{s \in S} P(s|p) I(\sum_i s_i = n) I(0 \in u' - s)}{\sum_{s \in S} P(s|p) I(\sum_i s_i = n)},$$

where S is the set of all possible sample tables, p is the population table (known to the data holder), sampling is simple random sampling without replacement, subtraction of tables is pointwise, and $I(X)$ is an indicator function, which returns 1 if X is true and 0 if it is false.

The strength of Smith and Elliot's (2005, 2008) method is that it can be applied with equal validity to unperturbed or perturbed tables, single or multiple table releases, as it is all based on the original base table manipulations. A second feature is that it is easily adapted to allow for the data snooper being less than 100% confident that they have recovered a true zero. It might be reasonable, for example, to consider that a snooper who was, say, 90% confident that a zero was a true zero, would be able to use this to make some probabilistic inferences. So, for any given sample table that a snooper might subtract from the published table, if all possible population tables contain a zero then we could say that the snooper will be 100% confident, if 90% of possible population tables do then we could say that they would be 90% confident. This second-level probability function allows in principle the method to take account of methods such as controlled tabular adjustment (see Chapter 4 for more discussion of this method) which does not produce tight bounds.

The SAP method is a promising way of understanding disclosure risk in tables of counts. Further work needs to be done with this technique to integrate it into data dissemination practice. Smith and Elliot (2005) identify computational issues in arriving at solutions for some problems. At least as defined by Smith and Elliot, the problem is NP-Complete.[5] This implies that some sort of non-exhaustive method like Markov Chain Monte Carlo (MCMC)[6] needs to be used in order to address large-scale problems.

Returning to a more conceptual question, just how important is negative attribution disclosure? Although an information-theoretic description implies that a negative attribution is a disclosure, this is not how it is commonly perceived. Stating that an arbitrary person is *not* a Ukrainian-born 25-year-old living in a flat in the UK city of Hull and working as a bank manager is simply not that interesting. On the other hand, knowing that an individual was not among those who passed a drug test is interesting. The weight of information obtained through an attribution—positive or negative—needs to be incorporated into the method in some way. Not all attributions are of equal import. In a general sense, this relates to the issue of sensitivity which we will now consider in detail.

3.5 What is at Risk?: Understanding Sensitivity

The disclosure risk assessment methods outlined above do not address the sensitivity of the information targeted by the data snooper. Implicitly they assume that if the attribution is successful then something important is disclosed about the identified population unit. That something is by definition not contained within the matching key. However, there are several other aspects of disclosed information that affect its *importance*:

- How easily could the information disclosed be obtained by other means?
- How damaging to the units concerned is the disclosure of that information?
- For publicized disclosures, how does the public at large perceive the damage from such a disclosure?

These three aspects are related. If you managed, for example, to identify the author Elliot in a data set and thereby uncover that he lives in a UK end-of-terrace house, this piece of information is not likely to cause sociopolitical earthquakes.

[5] In NP-Complete problems the size of the problem space grows exponentially as the size of the problem grows; this in turn leads to an increase in the computation time until the problem becomes intractable. For an overview of NP-Completeness see Garey and Johnson (1979).

[6] MCMC is a class of algorithms for sampling from probability distributions. Here we would be sampling the space of possible intruder knowledge states. See Robert and Cassella (2004) for an overview of MCMC.

The information is freely available to anyone who knows the author and sees him enter his house or indeed knows his address and engages in a few minutes of web searching. As it happens, the author is indifferent to anyone knowing what sort of house he lives in. Further, the general population is surely unlikely to see such information as damaging—"Elliot lives in end terrace!" will not sell many newspapers. So, although it might be a disclosure it is an uninteresting one; or to put it another way the information is not sensitive. Unfortunately, sensitivity is not homogeneous across any population. Some might be happy for their income to be public; others might regard that information as private. Even apparently mundane information could be sensitive in certain contexts. For example, a stalker might find the type of house their victim is living in useful in tracking them down. Generally, sensitivity of one variable is contingent on the values of other variables.

Based on a survey of the UK populations' perception of sensitivity, McCullagh (2007) found that the following types of information were regarded as sensitive by a majority of the population: personal contact details, financial data, data on race or ethnic origin, criminal record, biometric information, political opinions, membership in political organizations, clickstream data on a user's viewing of websites, religious or philosophical beliefs, genetic information, health information, sexual life information, educational qualifications, employment history, trade-union membership. The range of data types here does indicate the importance of the sensitivity issue, not least because many of these data are precisely those of interest to researchers.

One factor that McCullagh's analysis does not consider is that sensitivity may vary by value as well as by variable. For example, the information that someone does not have AIDS is probably not sensitive whereas the information that they do probably is. Also, being native to my country of residence I may regard country of birth as not particularly controversial, however those born outside their country of residence may in some cases regard their country of birth as sensitive.

Overall, sensitivity is a complex, multifaceted construct in which social, psychological, political, ethical, and legal issues all play roles. A key question for a DSO, but one beyond the scope of this book, is what proportion of a population needs to regard information sensitive before it must protect it from disclosure (the majority? 10%? just 1 person?). That said, some components of the construct are amenable to quantitative treatment. For example, differential-value sensitivity is often related to the population distribution of a variable as well as to variables assessing social meaning and value.

3.6 Summary

This chapter provides an overview of disclosure risk assessment techniques currently available for aggregate data and microdata. Assessment of disclosure risk is not an easy task because it is dependent on an ever-changing data environment, on data divergence (that is, the difference between the information stored in a data set about a given population unit and that which is available to a data snooper), and the consequences of a disclosure.

As we discussed in Section 3.1, when a disclosure risk measure is employed, thresholds are set below which the data can be released and above which more masking is necessary. In Section 3.2 risk metrics were developed to measure the extent to which a data snooper could use matching with an identified data file to compromise confidentiality. We developed risk measures both at the file level and at the record level. Our concern in Section 3.3 was this question: Given that a data set has been masked by a particular SDL method, how can we measure the disclosure risk? In this case the assumption of zero data divergence is untenable and leads to overestimation of the disclosure risk. If SDL was applied to a data product, we can attempt matching before and after the SDL to assess the impact of it on the data set to be released. In practice the matching is attempted through linkage of similar records in the external data set and the target data set. Another approach is to use the original data set, matching records in this unperturbed data set with records in the masked data set that was obtained from it. In Section 3.4 we considered aggregate data products. Finally, in Section 3.5 we considered the issue of assessing the sensitivity of the personal information that a data snooper might be able to infer from a data release.

Finally, what might be regarded as “sufficiently low risk” is a policy matter driven in large part by ethical considerations, not a statistical matter. In many cases such policy will be motivated by legislation and governmental regulations. Therefore, DSOs must determine acceptable levels of disclosure risk according to their own legal and ethical constraints. This will determine whether and to what extent SDL should be applied to the data before the dissemination takes place. It is to SDL techniques that we move in the next two chapters. We begin with techniques for data to be disseminated as tables.

Chapter 4
Protecting Tabular Data

Familiar to all of us, a statistical table displays aggregate information that is classified according to categories. Even in this age of electronic dissemination, tables remain important data products. In the past, DSOs published these tables in paper form as large statistical abstracts. Today, many DSOs provide users with an online capability for special tabulations—an easy way of generating their own tables. This table server mode of dissemination has been implemented as American *Fact-Finder* from the US Bureau of the Census, *Neighbourhood Statistics* from the UK Office for National Statistics, Multidimensional Statistical Database (BME) from the Brazilian Institute of Geography and Statistics, and *StatLine* from Statistics Netherlands, as well as other efforts of national statistical offices.

In many surveys, responses are categorical, and so tables of counts are natural for statistical reporting. Even for interval-scale data, DSOs often provide data products according to counts within banded quantities, such as the number of respondents with income levels from $30,000 to $40,000. Often these tables are simple counts of data units in a cross-classification, say income level by sex. But the table entries may also be weighted by selection probabilities, or the entries may be scaled up so that they can be interpreted as estimated counts of units in the total population.

As noted in Chapter 2, the occurrence of small counts in tables is usually understood to be potentially disclosive, because non-structural counts of zero may allow attributions about specific individuals known to be in the population. Additionally, low counts in a table are a problem whenever a data snooper has complete knowledge of a subset of the population contained within the table. In such a situation the snooper could remove or "subtract" those individual population units from the table possibly leaving residual non-structural zero counts.

Considerable effort has gone into developing SDL methods for tabular data (e.g., Cox et al., 2004; Duncan et al., 2001; Salazar, 2006a). Prominent among these methods is *cell suppression*, where the values of table cells that pose confidentiality problems are suppressed, as are values of additional cells that can be used to make inferences about confidential cells through the released table margins. Perturbation is also achieved through *controlled rounding*, *controlled tabular adjustment*, or *cyclic perturbation* techniques. All these approaches intend to replace the original cell values by what amounts to an interval of possible values, thus creating uncertainty about the original values for would-be snoopers. Indeed, no matter what

G.T. Duncan et al., *Statistical Confidentiality*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-7802-8_4,

the output looks like, a smart data snooper will start by computing the maximum and the minimum possible values for each cell, that is, the snooper will compute an interval. The wider the interval, the more *protected*[1] the cell, but also the larger the loss of information. For that reason, there are also other approaches (like partial cell suppression or interval publication) that directly replace cell values by intervals. The computational challenge associated with these approaches consists of finding an output which is *protected* but with a minimum of *information loss*. This challenge has been widely explored in recent years through techniques such as graph theory, mathematical programming, and artificial intelligence. This chapter aims to enable you to understand the principles and practice of the new procedures for tabular data that limit disclosure risk, even against sophisticated data snoopers, whilst still maintaining data utility. We first define the "properties" that we would like disclosure limited tabular output to satisfy, and then present SDL approaches that will provide outputs with these properties. In the first part we carefully define "protection" and "loss of information." In the second part we describe SDL approaches like cell suppression and controlled rounding. We answer three questions:

1. What is an appropriate definition of a disclosure-limited tabular output? (Section 4.1)
2. How can we design SDL mechanisms to provide such output using deterministic methods, especially through mathematical programming? (Section 4.2)
3. How can we design SDL mechanisms using stochastic methods, such as cyclic perturbation? (Section 4.3)

At their base, our answers to these questions are intuitively appealing and should contribute to the understanding of any reader of this book. On the other hand, in their implementation, our answers can be quite technical and involve mathematical formulations that will be of primary interest to that subset of readers who seek understanding at that level. In subsequent sections we will make clear when we make the transition to the more mathematical treatment.

A last remark before jumping inside the chapter: We are presenting the problem of protecting a table as a two-criterion optimization problem. One criterion is to maximize protection. The other criterion is to minimize the loss of information. As typically done when dealing with two-criterion problems, an effective way to approach this problem is to redesign it as a one-criterion problem where the most relevant criteria is limited through a constraint while the other is optimized. In our case, protection is prioritized and therefore it is taken as a limited resource that is seen as a constraint; thus the loss of information is the single criterion that should be minimized. In this chapter we will identify "loss of data utility" with "loss of information," although alternative formulations are possible.

[1] Here we will use the word "protection" to mean the width of this interval, so a *protected* cell is one where the interval width is sufficiently wide to be regarded as sufficiently disclosure limited.

4.1 Basic Concepts

In this section we present the basic concepts of SDL for tabular data. We want to be able to state when an output has good or bad properties, and to this end we develop these concepts: mathematical structure of a tabular array, the identification of risky cells,[2] the strategy of a data snooper to potentially disclosure information from a tabular data publication, the role of prior knowledge, inducing uncertainty in the data snooper, specifying upper and lower disclosure limits, identifying the extent of information loss, and the nature of disclosure auditing.

4.1.1 Structure of a Tabular Array

Tables aggregate one variable, usually called the *response variable*, in the original microdata according to classifications of specified categorical variables. For a study of the investment levels of enterprises, a DSO might construct a two-way table like Table 4.1 (adapted from Willenborg and de Waal (2001)). The three rows denote type of activity (I, II, and III) and the three columns denote geographical region (A, B, and C). For this table, investment is the aggregated variable while activities and regions are categorical variables. Traditionally, in order to be quickly useful, the simple table would be augmented with cells for the row totals, the column totals, and the grand total. In this example, the cell entries are not counts. Instead, each cell value represents the aggregation of the values of the response variable (e.g., capital investment of all the enterprises with the same activity and the same region).

Table 4.1 Total investment of enterprises by activity and region

	A	B	C	Total
Activity I	20	50	10	80
Activity II	8	19	**22**	49
Activity III	17	32	12	61
Total	45	101	44	190

Tables like Table 4.1 are called *magnitude tables.* A table can also be created by counting the number of records in each group, according to some explanatory variables. Tables of this type are called *count tables* (or *frequency tables*). A frequency table can be thought of as a special type of magnitude table in which the response variable is 1 if the unit falls in the cell and the aggregation is just the total. An important fact about a count table is that it always contains non-negative integer values. A magnitude table may have decimal or fractional values, or negative values, depending on the meaning of the response variable.

[2] In the literature you will often find references to “sensitive cells” used in the same way as we use “risky cells” here (i.e. cells which present a high disclosure risk). We have not adopted this term as it implies that the information contained within the cells is itself sensitive which it may not be.

Table 4.1 is two-dimensional. Many useful tables are higher-dimensional. For example, a table might be four-dimensional with attributes of activity, geographical region, type of ownership, and ranges of number of employees. Tables may have more complicated shapes. A *linked table* is one in which several tables may be combined through the linkage provided by some common cells. A *hierarchical table* contains marginal cells that group internal cells according to a hierarchical explanatory variable. For example, if a microdata set contains a variable representing the geographical location of each respondent, then a table with this variable as an explanatory variable can contain different marginal cells representing different groups of locations (wards, counties, local authorities, governments, regions, states, countries, etc.). Therefore, a hierarchical table has some cells that decompose into other smaller nested tables. More complicated versions are also possible, for example linked tables with hierarchical variables, like Table 4.2.

Table 4.2 Number of inhabitants (hypothetical data)

Unrounded data	Total	Male	Female	Young	Adult	Foreign-born	Native-born
North East	60, 593	29, 225	31, 368	13, 856	46, 737	34, 565	26, 028
North West	174, 414	78, 129	96, 285	25, 673	148, 741	3432	170, 982
Yorkshire and Humberside	108, 769	46, 119	62, 650	2342	106, 427	32, 223	76, 546
East Midlands	93, 346	43, 201	50, 145	23, 443	69, 903	23, 434	69, 912
West Midlands	131, 817	61, 046	70, 771	23, 878	107, 939	432	131, 385
East	107, 060	47, 376	59, 684	24, 532	82528	34, 233	72, 827
London	110, 811	49, 053	61, 758	17, 635	93, 176	3423	107, 388
South East	123, 359	50, 949	72, 410	34, 223	89, 136	4567	118, 792
South West	119, 863	44, 718	75, 145	35, 980	83, 883	56, 356	63, 507
England	1, 030, 032	449, 816	580, 216	201, 562	828, 470	192, 665	837, 367
Wales	95, 388	49, 579	45, 809	34, 989	60, 399	6454	88, 934
Scotland	124, 678	61, 327	63, 351	36, 789	87, 889	5643	119, 035
Great Britain	1, 250, 098	560, 722	689, 376	273, 340	976, 758	204, 762	1, 045, 336

No matter what the shape is, a statistical table is a collection of numbers—the internal cell values, the marginal cells, and so on—together with a collection of linear equations that specify the structural inter-relations of the cell values. Importantly, *a table satisfies a linear system of equations*. All tables in practice (including multi-dimensional, linked, hierarchical, and others) fit this structure. This fact is central to the mathematics of this chapter. We will preface the more mathematical material with a discussion, without equations, of the key concepts and implications for practice. The rest of this section provides the basic mathematical structure. It can be skipped by those interested in the overall ideas.

We denote the index set for the table cells by $I = \{1, \ldots, n\}$ and the index set for the equations by $J = \{1, \ldots, m\}$. The linear system of equations associated with a table with n cells linked by m equations is denoted by $\boldsymbol{My} = \boldsymbol{b}$, where $\boldsymbol{M}$ is a matrix with m rows and n columns, $\boldsymbol{y}$ is an n-dimensional vector of numbers, and $\boldsymbol{b}$ is an m-dimensional vector of numbers. In the standard case, each row of $\boldsymbol{M}$

contains many zeros, exactly one –1 and some +1s. The column index of the –1 corresponds to a marginal cell, and the column indices of the +1s correspond to the cells whose addition gives that marginal cell. The vector $\boldsymbol{b}$ is the zero vector in the standard case. Where more details are required, the linear system of equations will be represented by

$$\sum_{i \in I} m_{ij} y_i = b_j \quad \text{for} \quad j \in J,$$

where m_{ij} are the numbers in $\boldsymbol{M}$ and b_j the numbers in $\boldsymbol{b}$. As an example, Table 4.1 can be represented as an array of numbers for example:

$$\boldsymbol{a} = [20 \quad 50 \quad 10 \quad 80 \quad 8 \quad 19 \quad 22 \quad 49 \quad 17 \quad 32 \quad 12 \quad 61 \quad 45 \quad 101 \quad 44 \quad 190]$$

and the linear system this array satisfies is

$$y_1 + y_2 + y_3 - y_4 = 0$$

$$y_5 + y_6 + y_7 - y_8 = 0$$

$$y_9 + y_{10} + y_{11} - y_{12} = 0$$

$$y_{13} + y_{14} + y_{15} - y_{16} = 0$$

$$y_1 + y_5 + y_9 - y_{13} = 0$$

$$y_2 + y_6 + y_{10} - y_{14} = 0$$

$$y_3 + y_7 + y_{11} - y_{15} = 0$$

$$y_4 + y_8 + y_{12} - y_{16} = 0$$

For this example, $n = 16$ and $m = 8$. The rank of the matrix $\boldsymbol{M}$ is 7, and therefore it is possible to remove one of equation from the linear system without affecting the space of feasible solutions. For that reason, the number m of linear equations describing Table 4.1 may be reduced to coincide with the number of marginal cells ($m = 7$). Although this is the common situation in practice, there are tables where this does not occur. An example is Table 4.2, where initially $n = 91$ and $m = 53$, but the rank of the matrix $\boldsymbol{M}$ is strictly larger than 25 (the number of marginal cells in this table). Special SDL methodologies have been proposed (and implemented) for tables where the number of equations coincides with the number of marginal cells (see Salazar (2006b)). The methods described in this chapter apply to tables with any number of equations.

4.1.2 Risky Cells

In general, SDL protects information about an individual or an enterprise that is contained in the data to be published. When these data are in a magnitude table, the information needing protection could be the actual value contained in certain cells (as, for example, activity II and region C in Table 4.1). If a cell consists of just the contribution of a single respondent, then publishing this cell value will display the exact contribution of this respondent. The cells whose original values might reveal information on individual records are termed *risky cells*. This section shows some simple approaches to identifying the risky cells within a table. This task is referred to as the *Primary Problem* (Willenborg and de Waal, 2001). It consists of classifying each cell as risky or not, and is the first step: it determines which cells need disclosure limitation. The primary problem is typically addressed through heuristic approaches such as the dominance rule, prior/posterior ambiguity rule and *p* rule for magnitude tables, or the minimum frequency rule for count tables; we next describe some of these techniques.

4.1.2.1 Dominance Rule or (*n, k*)-Rule

A standard method for identifying risky cells in magnitude tables that have non-negative entries is to use the *dominance rule* or (*n*, *k*)*-rule* (Willenborg and de Waal, 2001). The rule depends on two parameters, a positive integer *n* and a percentage *k*. The DSO adjusts *k* and *n* to obtain the desired stringency. In application, the respondents grouped into each cell in the table are sorted by decreasing order of their response values. For each cell, if the largest *n* respondents in the list contribute at least *k*% of the total value of this cell, then the cell is classified as *risky*. For example, if there are five records contributing values of 55, 32, 16, 5, and 4, respectively, then the cell showing the corresponding total 112 would be considered risky if, for example, *n* had been set to 3 and *k* had been set to 75. This rule is motivated by the fact that the *n* major respondents can guess with sufficient precision the value of the other respondents to the same cells. Some DSOs do not publish the *n* and *k* values that they use. If these values are published, then one should account for this when solving the secondary problem (see Section 4.1.4). As we shall see later, some SDL techniques (such as cell suppression) protect the risky cell without taking into account the rule and the parameters used in the primary problem.

4.1.2.2 Prior/Posterior Ambiguity Rule

Under the *prior/posterior ambiguity rule* a cell is deemed risky when another contributor could reasonably approximate the value of the largest contributor within that cell. It is also called the *p/q rule* as it needs two parameters *p* and *q* (with $p<q$) to be specified. It requires publicly available information to provide an estimate (prior to publication of the table) of the contribution of one contributor within *q*%. After the publication of the table, this estimate would be within *p*%. In the *p/q* rule the ratio *p*/*q* can be thought of as the information gain through publication of the table. If the information gain is unacceptable, the cell is declared *risky*. The DSO specifies the

values p and q to yield an acceptably low level of information gain. Using the above example, if $p = 25$ and $q = 50$, then the cell is risky because 112 is bigger than 2 times 55.

4.1.2.3 *n*-Rule

With a frequency table the dominance and prior/posterior ambiguity rule are not relevant. Many DSOs implement the n-rule, effectively picking an arbitrary small value (say, n smaller or equal to 3) and denoting counts of that size or less as risky. Following the subtraction principle outlined in Chapter 3, this rule has some obvious merit: the smaller n is the less a priori information a snooper will need in order to derive a count of zero.

A variety of other heuristic rules have been proposed in the literature, see for example Loeve (2001), Robertson and Ethier (2002), and Domingo-Ferrer and Torra (2004). All of them are based on common sense, and it is always possible to argue about the advantages and disadvantages of each rule when applied to a particular table. Therefore, the choice of a particular rule for declaring a cell risky depends on the context in which the table is developed. In the next section we put aside the issue of how risky cells are specified and assume that some well-formulated method has been employed.

4.1.3 The Secondary Problem: The Data Snooper's Knowledge

Once risky cells have been identified in a table (i.e., after the Primary Problem is solved), the DSO must choose an SDL approach to protect the data. This task is called the *Secondary Problem.*

4.1.3.1 A Priori Knowledge

DSOs prudently assume that a data snooper has some knowledge about a cell value even before it is published. But depending on the type of knowledge, it may be possible to develop a technique to protect a table against a data snooper with this knowledge. One situation where techniques exist is when the knowledge consists of an interval of possible values for the cell. For example, in a magnitude table if a data snooper was the contributor to cell i and his contribution was five units, then the data snooper is sure that the cell value is at least five units (i.e., $y_i \geq 5$), regardless of any published material. In this example, the interval of this data snooper for cell i is $[5, U]$ assuming s/he also knows that a value over U units is impossible. Of course, different data snoopers may have different knowledge (i.e., different intervals). For each snooper k and each cell i, the interval is named *a priori knowledge* and is denoted by $[\mathrm{lb}_i^k, \ \mathrm{ub}_i^k]$. For a simple illustration of a data snooper's a priori knowledge, suppose that each snooper k knows a lower bound lb_i^k and an upper bound ub_i^k on the potential value y_i of each cell i, i.e.,

$$\mathrm{lb}_i^k \leq y_i \leq \mathrm{ub}_i^k.$$

These two bounds are called *external bounds*. Since each data snooper has this knowledge then the DSO must take it into account when protecting the data. See Smith and Elliot (2008) for an approach to modeling of different levels of knowledge that a data snooper might have (also outlined in Chapter 3).

There are some cases where a snooper knows external bounds and the DSO knows what those bounds are. The obvious case is where the response variable is structurally non-negative. In this case, an outsider with no other information can be associated with the external bounds $\mathrm{lb}_i^k = 0$ and $\mathrm{ub}_i^k = +\infty$. Another situation is when the data snooper knows a range of values within the given percentages around the true value (for example, lb_i^k is 50% of the cell value a_i, and ub_i^k is 150% of the cell value a_i). This might be reasonable when similar information has been published in previous years in similar regions, and so on.

If the DSO wants to take into account more elaborate hypotheses on a data snooper's a priori knowledge, then protecting a table can be more complex. For example, suppose a data snooper knows that a given cell has a value that conforms to a particular probability distribution. Then, the DSO cannot pretend to find a publication such that minimum and maximum values computed by the data snooper should define a wide-enough interval. Instead, the DSO can only affirm protection with a probability level smaller than one. In other words, protecting a table against snoopers with this type of probabilistic a priori knowledge is more complicated than protecting a table against snoopers with only knowledge based on deterministic external bounds. While in this second situation (deterministic knowledge) there exist widely used techniques that ensure (mathematically provable) protection, in the first situation (probabilistic knowledge) there are still many open problems. Further compounding this issue is how the DSO can know the probability distribution that the snooper is using. Any choice of probability distribution is therefore arbitrary, and consequently DSOs have tended to use the operational assumption that data snoopers have deterministic external bounds and it is this approach that we focus on in this chapter.

4.1.3.2 The Output Pattern

The intent of SDL in the context of tabular aggregates is to transform the given table before publication in order to induce "uncertainty" about target values on the part of a data snooper. "Uncertainty" here means that, from the publication and also possibly from the a priori knowledge, a data snooper interested in a risky cell cannot infer a close approximation of its original value (i.e., the minimum and maximum possible values that the data snooper will compute using his knowledge determine a wide-enough interval). To this end, a publication should be seen by a data user (whether legitimate or a snooper) as just a collection of possible original (or source) tables. One of these tables will be the original one, containing the exact values for the risky cells but, if effective, the disclosure limitation will guarantee that there are also other potential tables from the data snooper's point of view. Following Fischetti and Salazar (1998) we will refer to the published information as an *output pattern*.

This terminology emphasizes that, although a publication may look like a single table, it actually represents a set of tables that is all of the possible tables which might have lead to the publication and any one of which from the snooper's point of view might be the original one. We assume that the aim of the SDL process is to guarantee that the number of possible tables is neither too small (i.e., there would be too much residual disclosure risk) nor too large (indicating too little data utility).

Section 4.2 illustrates the set of tables associated with four types of output patterns, each one associated with a particular SDL method that could be employed to solve the secondary problem.

The following material through the end of Section 4.1.5 is more mathematical. The key idea is that the a priori knowledge of the data snooper can often be expressed as upper and lower bounds on cell entries of the table. It is easy to incorporate such information as constraints in a mathematical programming formulation for what the publication says about the original or source table. In principle this formulation is an integer linear programming (ILP) problem, because the cell entries are integers and the constraints are inequalities about linear combinations of the cell entries. ILP problems are computationally difficult to solve, but in this context the corresponding linear programming (LP) problem is both easy to solve and often provides a solution that is an adequate approximation to the solution of the ILP problem.

When the data snooper is taken to have a priori knowledge through external bounds, the candidate tables can be specified through a set of linear constraints, containing the equations describing the table structure and the bounds known by the snooper. To be more precise, a table $\boldsymbol{y} = [y_i : i \in I]$ is a candidate to be the original table for the data snooper k if it satisfies

$$\sum_{i\in I} m_{ij} y_i = b_j \quad \text{for all} \quad j \in J, \tag{4.1}$$

$$\text{lb}_i^k \leq y_i \leq \text{ub}_i^k \quad \text{for all} \quad i \in I. \tag{4.2}$$

Equations (4.1) ensure that the values are congruent with the table structure and inequalities (4.2) ensure that the values are consistent with the a priori knowledge. When other information on a generic candidate table $\boldsymbol{y}$ is taken to be known by a data snooper, this information should be added to the above system in a mathematical style (i.e., writing constraints). In principle, these constraints may be probabilistic or non-linear information, and so difficult to manage. In practice, however, DSOs rarely know what an arbitrary data snooper might know beyond the external bounds, and in most cases DSOs work assuming that the data snooper's prior knowledge only consists of external bounds (i.e., that the value for cell i is between lb_i^k and ub_i^k).

A table which satisfies the system (4.1) and (4.2) is called a *congruent table*. Obviously, the source table (say $\boldsymbol{a} = [a_i : i \in I]$) is always a congruent table, but the key point of an SDL method is to guarantee that other congruent tables do also exist. Still, not all the solutions of (4.1) and (4.2) are candidate tables for data snooper k because constraints related to the disclosure limitation method to be employed are missing. For example, if the applied SDL method was rounding, then some solutions of (4.1) and (4.2) are infeasible. Section 4.2 gives those constraints for each of four different SDL methods. These missing constraints are modeled through linear inequalities in many cases, and therefore the set of candidate tables for a data snooper k can be represented by a *polyhedron* in an n-dimensional space. If we assume that the data snooper's a priori information includes the knowledge that the cell values are integers, then a possible original or source table—from the data snooper's viewpoint—corresponds to an integer vector contained within this polyhedron. Let S^k be the set of all the candidate tables which are congruent with the table structure for a given output pattern, and the a priori knowledge of the data snooper k. Remember: what is important is that this set S^k should not be too small (which implies too much disclosure risk) nor too large (which implies too little data utility).

As already mentioned, from a given output pattern, each data snooper will try to compute the smallest and largest possible value for each risky cell. This can be done by solving two optimization problems (hereafter "the data snooper problems") in conjunction:

$$\underline{y_p^k} := \text{ minimize } y_p \text{ subject to } \boldsymbol{y} \in S^k, \tag{4.3}$$

$$\overline{y_p^k} := \text{ maximize } y_p \text{ subject to } \boldsymbol{y} \in S^k. \tag{4.4}$$

If all the constraints defining S^k are linear constraints, then the data snooper problems are two standard LP problems, which can be easily solved using readily available software. This is the situation when we are protecting a magnitude table where the response variable may assume fractional values and the a priori knowledge of the snooper consists of external bounds. If the only non-linear constraint is the fact that the cell values are integers (which is the case when protecting a frequency table), then solving the data snooper problem requires ILP tools. With modern computing, it is easy for a data snooper to solve large LP problems, and also some moderate-sized ILP problems. For simplicity in the following we will assume that the data snooper problems can be solved by using LP techniques (the basic ideas can be adapted to work also with ILP).

After solving the data snooper problems, data snooper k will have calculated the "uncertainty" mentioned above as the interval $[\underline{y_p^k},\ \overline{y_p^k}]$ of potential values for each cell p. Note that for each number $y_p \in [\underline{y_p^k}, \overline{y_p^k}]$ there should exist a potential table $\boldsymbol{y} \in S^k$ which has number y_p in cell p. Now we are ready to define the situation where an output pattern sufficiently limits the disclosure risk for the information a_p in cell p against data snooper k.

4.1.4 Disclosure Limitation

We assume that the DSO sets up two parameters LL_p^k and UL_p^k for each data snooper k and risky cell p. These are called *lower* and *upper disclosure limitation levels*, respectively. In this approach, an output pattern is deemed to *protect* cell p against data snooper k if both of these conditions hold:

$$\overline{y_p^k} - a_p \geq \mathrm{UL}_p^k \tag{4.5}$$

$$a_p - \underline{y_p^k} \geq \mathrm{LL}_p^k. \tag{4.6}$$

These two conditions guarantee that the lower and the upper bounds that snooper k will compute, respectively, are far enough from the original value a_p. It may be of interest to some DSOs to also add other requirements to a disclosure-limited output, like for example

$$\overline{y_p^k} - \underline{y_p^k} \geq \mathrm{SL}_p^k \tag{4.7}$$

$$\overline{y_p^k} + \underline{y_p^k} \geq \mathrm{CL}_p^k. \tag{4.8}$$

where the levels SL_p^k and CL_p^k are other two given input parameters. The first parameter ensures a minimum difference between the two values $\underline{y_p^k}$ and $\overline{y_p^k}$ that the data snooper will compute. It could be useful when one is interested in protecting only the exact value a_p, in which case $\mathrm{UL}_p^k = \mathrm{LL}_p^k = 0$ and $\mathrm{SL}_p^k = 1$. The second parameter is useful when one wants to ensure that the median point of the interval $[\underline{y_p^k}, \overline{y_p^k}]$ that the snooper will compute should be far from the original value a_p. For example, if $\mathrm{CL}_p^k = 3a_p/2$, then the median point will be greater or equal to $3a_p/4$. For simplicity we assume that the DSO only wants to impose conditions using LL_p^k and UL_p^k.

As indicated, the disclosure limitation levels are set by the DSO. For example, simple values for the upper and lower disclosure limitation levels are percentages of the true value of the cell. In more sophisticated situations where the data snooper k is one of the original respondents, the disclosure limitation levels could be chosen to be proportional to the respondent's contributions s_p^k to the nominal value a_p of the cell p and/or to the complement $a_p - s_p^k$ (Cox, 1981; Robertson, 2000; Sande, 1984). An elementary requirement is that

$$\mathrm{lb}_p^k \leq a_p - \mathrm{LL}_p^k \leq a_p \leq a_p + \mathrm{UL}_p^k \leq \mathrm{ub}_p^k$$

for each data snooper k and each risky cell p.

4.1.5 Loss of Information

As described above, after disclosure limitation has been applied the published information could be seen as a set of possible tables. All the tables in this set are coherent with the table structure and the a priori information, and among this collection there is the original table $\boldsymbol{a}$. As the size of the set increases, the data utility of the output pattern can be assumed to decrease (because the uncertainty of cell values increases). In other words, an output leading to a larger set of potential tables is more disclosure-limited but has smaller data utility.

The *loss of information* of an output pattern is related to the number of potential tables in the set S^k for each user k of the output pattern. When the only table in S^k is $\boldsymbol{a}$, the loss of information is zero. Otherwise, the loss is positive.

In general, it is difficult to calculate the number of potential tables in a collection S^k. When the a priori knowledge can be expressed as linear inequalities, as cited above, the collection S^k can be represented as a polyhedron in the n-dimensional real space, and therefore the loss of information should be related to its volume. However, even computing this volume is a complicated task, and therefore the *loss of information* of a pattern is typically replaced by a simpler function which depends on the method that produced the pattern. Section 4.2 will illustrate four SDL methods, and for each one the definition of the "loss of information" for a given output pattern will be given.

4.1.6 The DSO's Problem

As we have noted before, when protecting data the DSO must solve two problems. One is the primary problem and concerns detecting when a cell is risky. The other is the secondary problem and aims at protecting the risky cells. When dealing with tabular data the primary problem is easily tackled with simple techniques. The secondary problem, however, is complex. It can be seen as the bi-criteria optimization problem of finding an output pattern (i.e., a collection of tables coherent with the structure of the table and with the a priori knowledge of each data snooper) such that it has the minimum loss of information (highest data utility) and the maximum confidentiality. In other words, we want to find a pattern that produces a set S^k of potential tables (for snooper k) with a large volume to guarantee protection and with a small volume to guarantee utility. Since the two criteria are opposed, the standard approach to overcome this difficulty is to make one criterion a constraint and leave the other criteria in the objective function. As legal and ethical requirements arguably compel the DSO to view confidentiality as critical, this is usually the criterion that is controlled through constraints (i.e., bounded by disclosure limitation levels which are deemed to meet the legal and ethical requirements). Given this the secondary problem becomes: amongst the set of disclosure-limited outputs which satisfy these constraints, find the one with minimum loss of information. In other words, agencies typically view the disclosure risk as a constraint and the data utility as an objective.

Solving this reformulated problem may still be complicated due to the size of the table to be protected; therefore the DSO may relax the requirement of finding an optimal pattern. Accepting a non-optimal pattern in the sense that there may exist a better pattern according to the loss of information criteria. However, what cannot be relaxed are the protection constraints, which means that the pattern must be sufficiently disclosure-limited, no matter if it is optimal or sub-optimal, according to information loss.

4.1.7 Disclosure Auditing

Due to the complexity of generating disclosure-limited output patterns, some researchers in SDL (see, for example, Doyle et al. (2001)) have proposed techniques to find "good" patterns with no inherent guarantee (under the described assumptions) of the disclosure limitation level requirements, and which therefore do not necessarily provide "disclosure-limited data" as we define it here. Therefore, it is necessary to check the patterns before data are disseminated. To this end, the agency must solve the two data snooper problems mentioned in Section 4.1.5 to compute $\underline{y_p^k}$ and $\overline{y_p^k}$ for each risky cell p and each snooper k, and compare these values with the required disclosure limitation levels. This is often called the *Disclosure Auditing Phase*. If the audit suggests that the output is not disclosure-limited, then the agency must change the technique to produce a different output. This process will then need to be iterated until an output pattern which meets the DSO's disclosure limitation criteria is found. Note that this disclosure auditing phase is only necessary when the implemented approach does not guarantee protection, which is not the case with the methods described in this chapter. The methods described in this chapter guarantee protection (mathematically provable) under the discussed assumption (for example, snooper's a priori information consists in external bounds and not in probabilistic knowledge). In the DSO context, solving the *auditing problem* is doing the snooper's work, which means computing the minimum and the maximum value for each risky cell before releasing the output. Then the interval computed for each risky cell must conform to the specified disclosure limitation levels, and the output can only be released if all the disclosure limitation levels hold. Hence, auditing aims at checking that in the worst situation (given our assumptions) the table is protected. Of course, if a data snooper has information that is outside the scope of our assumptions, then the output may not be protected.

4.2 Four Methods to Protect Tables

This section describes four approaches to protecting tables: cell suppression, interval publication, controlled rounding, and cell perturbation. Each approach generates a particular DSO problem. The key to operationalizing each of these approaches is to write appropriate *optimization models* for the problem, and to use mathematical programming techniques to develop algorithms and find optimal or near-optimal solutions. The pioneers in using mathematical programming tools in SDL are Cox (1980) and Sande (1984). See also Zayatz (1992) and Fischetti and Salazar (1998).

4.2.1 Cell Suppression

Cell suppression is the most popular technique for protecting confidentiality in statistical tables. The standard cell suppression technique protects the risky information by hiding (*suppressing*) the values of some cells. The suppressed cell value is usually replaced with some symbol, for example, an asterisk or an "s." To begin cell suppression, the set of risky cells P are suppressed; they are called *primary suppressions*. However, since marginal cell values are often released and are obtained by adding internal cell values in the table, simply suppressing primary cell values typically will not limit disclosure. This is clear in Table 4.1 where there is no limitation of disclosure if we only suppress the risky cell (cell in activity II and region C). So, to ensure sufficient disclosure limitation other cells must also be suppressed. These cells must be computed (they will be the solution of the secondary problem) and are called *secondary suppressions* or *complementary suppressions*. Here, the optimization problem is to determine the secondary suppressions, that is, which other cells should be suppressed so that the risky cell values cannot be recovered by subtraction from the marginal totals.

The problem of finding an acceptable output pattern with minimum loss of information is a difficult combinatorial optimization problem known as the *Cell Suppression Problem* (hereafter: CSP); see for example Willenborg and de Waal (2001) for further discussion. The task is so complex that in the literature there are mainly heuristic algorithms (i.e., procedures providing suppression patterns that are not necessarily optimal—and indeed probably overly disclosure limited) for many situations. For example, a relevant situation occurs when there is an entity which contributes to several cells, leading to the *common respondent problem*. Possible simplifications to approach this situation consist in replacing several data snoopers by one "super" data snooper with "disclosure limitation capacities" (see, for example, Jewett (1993) or Sande (1999) for details), or on aggregating some risky cells into new "union" cells with stronger disclosure limitation level requirements (see, for example, Robertson (2000)). But even the simpler problem that considers just one data snooper is classified in computational complexity theory as a *strongly NP-hard problem.* For practical purposes this means that an algorithm cannot be found for the exact solution of CSP that guarantees efficient performance for all inputs (see, for example, Kelly et al. (1992)). Previous work concentrates on solving the CSP for a two-dimensional table together with its marginal totals and considering only a single data snooper. Heuristic solution procedures have been proposed by several authors, including Cox (1980, 1995), Sande (1984), Kelly et al. (1992), and

Carvalho et al. (1994). Kelly (1990) proposed a mixed ILP formulation involving a huge number of variables and constraints (for instance, the formulation involves more than 20,000,000 variables and 30,000,000 constraints for a two-dimensional table with 100 rows, 100 columns, and 5% risky entries). Geurts (1992) refined this model, and reported computational results by dealing with small-size instances, the largest instance solved to optimality being a table with 20 rows, 6 columns, and 17 risky cells. Heuristics for three-dimensional tables have been proposed by Robertson (1994), Sande (1984), and Dellaert and Luijten (1999). Fischetti and Salazar (1999) proposed a new method capable of solving to proven optimality, on a personal computer, two-dimensional tables with about 250,000 cells and 10,000 risky entries. An extension of this method capable of solving to proven optimality realistic three- and four-dimensional tables that protect against one data snooper is presented in Fischetti and Salazar (2000). These results can be extended to the case of multiple data snoopers. Salazar (2008) provides technical details.

We flag the remaining material of this section as requiring more mathematics. The key idea is that the problem of what cells to suppress can be expressed as a mathematical programming problem, specifically as an ILP problem.

An output pattern in cell suppression is then defined by the cells to be suppressed (hereafter SUP), both the primary suppressions P and the secondary suppressions. Since P is a subset of SUP, the optimization problem is to determine the other cells in SUP. The feasible region S^k for the problems associated to data snooper k is defined by

$$\sum_{i \in I} m_{ij} y_i = b_j \quad \text{for all} \quad j \in J,$$

$$y_i = a_i \quad \text{for all} \quad i \in I \backslash \text{SUP},$$

$$\text{lb}_i^k \le y_i \le \text{ub}_i^k \quad \text{for all} \quad i \in \text{SUP}.$$

Table 4.3 illustrates a pattern for the instance in Table 4.1, where SUP is the set of cells containing an asterisk. Assuming that there is one data snooper who just knows that each missing value is a non-negative number (i.e., $\text{lb}_i^k = 0$ and $\text{ub}_i^k = +\infty$), then the minimum value $\underline{y_{\text{II,C}}}$ for the risky cell in row Activity II and column C can

Table 4.3 Cell suppression pattern for Table 4.1

	A	B	C	Total
Activity I	20	50	10	80
Activity II	$*$	19	$*$	49
Activity III	$*$	32	$*$	61
Total	45	101	44	190

be computed by solving an LP problem in which the values $y_{i,j}$ for the suppressed cells in row i and column j are treated as optimization variables, namely

$$\underline{y_{\mathrm{II,C}}} := \min y_{\mathrm{II,C}}$$

subject to

$$y_{\mathrm{II,A}} + y_{\mathrm{II,C}} = 30$$

$$y_{\mathrm{III,A}} + y_{\mathrm{III,C}} = 29$$

$$y_{\mathrm{II,A}} + y_{\mathrm{III,A}} = 25$$

$$y_{\mathrm{II,C}} + y_{\mathrm{III,C}} = 34$$

$$y_{\mathrm{II,A}} \geq 0,\ y_{\mathrm{III,A}} \geq 0,\ y_{\mathrm{II,C}} \geq 0,\ y_{\mathrm{III,C}} \geq 0.$$

Note that the right-hand-side values are known to the data snooper, as they can be obtained as the difference between the marginal and the published values in a row/column.

The maximum value $\overline{y_{\mathrm{II,C}}}$ for the risky cell can be computed in an analogous way, by solving the LP problem maximizing $y_{\mathrm{II,C}}$ subject to the same constraints. Given the a priori knowledge of the external bounds (non-negativity on this example), each solution is a table congruent with the published suppression pattern in Table 4.3. Since $\underline{y_{\mathrm{II,C}}} = 5$ and $\overline{y_{\mathrm{II,C}}} = 30$, the risky information is "disclosure-limited" being within the *disclosure limitation interval* [5, 30]. If this interval is considered sufficiently wide by the DSO according to its confidentiality requirements, then the pattern in Table 4.3 is *feasible* or *valid* (i.e., acceptable); otherwise, different or further secondary suppressions will be needed. Remember that the confidentiality requirements are fixed by the DSO through disclosure limitation levels (like LL_p^k, UL_p^k, SL_p^k and/or CL_p^k for each cell p and each snooper k).

In order to define the "loss of information" of a cell suppression pattern, we need a measure, denoted w_i, of the loss of information when cell i (with actual value a_i) is not published. Typical definitions of "loss of information" (see, for example, Willenborg and de Waal (2001)) define w_i according to one of the following options:

1. a_i,
2. 1,
3. $\log(a_i)$,
4. the number of responses in the microdata contributing to value a_i of cell i,
5. a (linear) combination of the above criteria.

Then the loss of information of an output pattern determined by SUP is defined as sum of w_i for all $i\in$SUP. Salazar (2010) details the experience of writing this model as simple computer code (for both magnitude and count tables).

4.2.2 Interval Publication

In *Interval Publication* a publication pattern is a set of intervals, one $[y_i^-,\ y_i^+]$ for each cell i. For each cell i and each value $y_i \in [y_i^-, y_i^+]$ there is a congruent table $\boldsymbol{y}'$ where $y'_i = y_i$ and $y'_l \in \left[y_l^-, y_l^+\right]$ for all $l = I$. For Table 4.1 a feasible pattern could be that in Table 4.4.

Table 4.4 Interval publication pattern for Table 4.1

	A	B	C	Total
Activity I	[18 . . . 24]	50	[6 . . . 12]	80
Activity II	[4 . . . 10]	19	[20 . . . 26]	49
Activity III	17	32	12	61
Total	45	101	44	190

Under the interval publication technique, the feasible region S^k for the data snooper problems associated with data snooper k is defined by

$$\sum_{i\in I} m_{ij} y_i = b_j \quad \text{for all} \quad j \in J,$$

$$y_i^- \le y_i \le y_i^+ \quad \text{for all} \quad i \in I,$$

$$\text{lb}_i^k \le y_i \le \text{ub}_i^k \quad \text{for all} \quad i \in I.$$

The input parameters w_i^+ and w_i^- can be defined by common-sense measures similar to those discussed for the cell suppression technique.

The loss of information in publishing $[y_i^-, y_i^+]$ instead of a_i can be measured as a proportion of $a_i - y_i^-$ and of $y_i^+ - a_i$. To this end, two input parameters w_i^+ and w_i^- are given for each cell i, and the “loss of information” of a pattern is defined as

$$\sum_{i\in I} w_i^+(y_i^+ - a_i) + w_i^-(a_i - y_i^-).$$

The interval publication method was introduced by Fischetti and Salazar (2003) against one data snooper with the name the *partial cell suppression method*. The method shares features with the cell suppression method. Indeed, from a cell suppression pattern a data snooper can replace the missing values with intervals of possible values. Therefore, from a data snooper’s point of view, patterns produced by the two methods can be considered equivalent. Nevertheless, the cell suppression method is a dichotomous approach, where a cell must be published or not, whilst with the interval publication method the value of each risky cell is replaced by an interval that can vary in width. That is why the interval publication method was originally called *partial* cell suppression. This extra freedom in interval publication has the advantage of providing patterns containing a smaller number of congruent tables than the ones from Cell Suppression, and hence increasing the data utility of the publication pattern. In other words, the set of congruent tables associated with a valid cell suppression pattern coincides with the set of congruent tables associated with a valid interval publication pattern, but the reverse is not true. Since the region of valid patterns under interval publication contains the region of valid patterns in cell suppression, it is possible to find solutions with smaller loss of information.

Another important advantage of the interval publication method is that the DSO problem associated to it (called *Interval Publication Problem*, or IPP in short) is much simpler computationally. Nevertheless, optimal interval publication patterns have the disadvantage of containing intervals on more cells than there are missing values in optimal cell suppression patterns. For technical details see Salazar (2008).

4.2.3 Controlled Rounding

Cell suppression and interval publication both can be seen as revealing possible values for the cells of the table. However, in some situations, the DSO may prefer to publish only one value for each cell. In doing so, the DSO is not misinforming the data user, because if a value different than the original value is published, the DSO will inform the data user that the published value may not be the original one in a cell, but rather a "close" value. The data user could then choose to replace the published value by an interval of potential values, thus being in a similar situation to that generated by the previous methods. The difference is that, by using such a method, the set of intervals differ, and therefore depending on the nature of the source table the data utility of the output pattern may be better (or worse).

With *controlled rounding,* we specify a base number r_i for each cell i. Note that r_i may be a constant independent on the cell i (as traditionally used; for example, $r_i = 5$ for all i) or may be a different value for each cell. In other words, base numbers are often identical across the whole of a table release, though this is not a defining feature of the technique. What is released is a table with each entry rounded to the specified base. Tables 4.5 and 4.6 give examples of patterns when $r_i := 5 (i \in I)$ for the instances in Tables 4.1 and 4.2, respectively. The method is called controlled because of a constraint to have the sum of the published entries in each row and column equal to the appropriate published marginal totals. Mathematical programming methods are used to identify a controlled rounding pattern for a table. Indeed, the optimization problem of matrix rounding was already investigated the early article of Bacharach (1966).

Table 4.5 Controlled rounding pattern from Table 4.1 with $r_i = 5$ for all $i \in I$

	A	B	C	Total
Activity I	20	50	10	80
Activity II	10	20	20	50
Activity III	15	30	15	60
Total	45	100	45	190

Table 4.6 Controlled rounding pattern from Table 4.2 with $r_i = 5$ for all $i \in I$

Rounded data ($r_i = 5$)	Total	Male	Female	Young	Adult	Foreign-born	Native-born
North East	60,595	29,225	31,370	13,855	46,740	34,565	26,030
North West	174,415	78,130	96,285	25,675	148,740	3430	170,985
Yorkshire and Humberside	108,770	46,120	62,650	2340	106,430	32,225	76,545
East Midlands	93,345	43,200	50,145	23,445	69,900	23,435	69,910
West Midlands	131,815	61,045	70,770	23,875	107,940	430	131,385
East	107,060	47,375	59,685	24,530	82,530	34,235	72,825
London	110,810	49,055	61,755	17,635	93,175	3420	107,390
South East	123,360	50,950	72,410	34,225	89,135	4570	118,790
South West	119,860	44,715	75,145	35,980	83,880	56,355	63,505
England	1,030,030	449,815	580,215	201,560	828,470	192,665	837,365
Wales	95,390	49,580	45,810	34,990	60,400	6455	88,935
Scotland	124,675	61,325	63,350	36790	87,885	5640	119,035
Great Britain	1,250,095	560,720	689,375	273,340	976,755	204,760	1,045,335

What follows next is a mathematical formulation for an implementation of controlled rounding.

Let $\lfloor a_i \rfloor$ be the multiple of r_i obtained by rounding down a_i, and $\lceil a_i \rceil$ be the multiple of r_i obtained by rounding up. When r_i is such that $\lfloor a_i \rfloor = \lceil a_i \rceil$ then we redefine $r_i = 0$. A pattern in the controlled rounding method is a congruent table $\boldsymbol{v} = [v_i : i \in I]$ such that

$$v_i \in \{\lfloor a_i \rfloor, \lceil a_i \rceil\} \tag{4.9}$$

The values r_i are known by the data snoopers. The feasible region S^k for the data snooper problems associated with data snooper k is defined by

$$\sum_{i \in I} m_{ij} y_i = b_j \quad \text{for all} \quad j \in I,$$

$$v_i - r_i \leq y_i \leq v_i + r_i \quad \text{for all} \quad i \in I,$$

$$\text{lb}_i^k \leq y_i \leq \text{ub}_i^k \quad \text{for all} \quad i \in I.$$

The "loss of information" of a cell can be defined as the absolute difference between the nominal value and the published value, and then the loss of information of a pattern is the sum of all the individual loss of information values.

The main difficulty with this method is that a feasible output pattern does not always exist. A necessary condition for feasibility is that

$$\mathrm{LL}_p^k + \mathrm{UL}_p^k \le 2r_p \quad \text{for all} \quad k \in K \text{ and } p \in P.$$

But this condition is not sufficient (even if all disclosure limitation levels are zero). The combinatorial problem of finding a disclosure-limited pattern with minimum information loss is called the *Controlled Rounding Problem* (CRP). The problem was first introduced by Bacharach (1966) in the context of replacing non-integers with integers in tabular arrays, and arises in several application contexts. To reduce the complexity of finding a feasible pattern, typically a uniform base number is used, as in the above example. However, although there always exists a feasible pattern for a two-dimensional table, there is no such guarantee for multi-dimensional tables with marginal totals. Causey et al. (1985) showed the simple infeasible 2 × 2 × 2 instance illustrated in Table 4.7. Sub-table (c) is the sum of sub-tables (a) and (b). Using one as base number, it is easy to check that no solution is possible without changing an original integer value (see, for example, that cell in row 1, column 1, sub-table (a) cannot be fixed to 0 nor to 1).

Tables 4.7(a–c) Table without a rounding solution if the integer values should remain unchanged

0.5	0	0.5
0	0	0
0.5	0	0.5

(a)

0	0.5	0.5
0.5	0	0.5
0.5	0.5	1

(b)

0.5	0.5	1
0.5	0	0.5
1	0.5	1.5

(c)

Kelly et al. (1990) proposed a branch-and-bound procedure for the case of three-dimensional tables, based on the LP relaxation of an ILP problem. This type of technique is commonly used when solving models in mathematical programming with binary variables. The idea is to partially enumerate all possibilities by choosing a variable to be fixed in a sub-problem to value 0 and in another sub-problem to value 1. This splitting operation is called *branching*. To avoid full enumeration when finding a solution with the minimum loss of information, a lower estimation of the loss of information at each sub-problem is computed. This estimation is called the *bound*, and it avoids the necessity of a long sequence of branching operations.

Heuristic methods for solving CRP on multi-dimensional tables have been proposed by several authors to find near-optimal solutions. Kelly et al. (1990, 1993) and Fischetti and Salazar (1998) proposed branch-and-bound procedures for its resolution based on computing the bound through solving an LP problem. See Salazar et al. (2004) and Salazar (2006a) for details of a general model for linked and hierarchical tables against different data snoopers.

Finally, the described method is typically known as *zero-restricted* controlled rounding because each cell value has at most two options (i.e., either $\lfloor a_i \rfloor$ or $\lceil a_i \rceil$). This strong constraint is the reason why some instances may be infeasible. To relax this constraint and possibly find a pattern to be published, Salazar (2006a) has proposed a more general variant where the DSO must set not only the base numbers but also another parameter γ. This parameter γ is a non-negative integer number that specifies the maximum number of roundings that can be applied either up or down for each cell value, that is, a_i must be chosen inside $\{\lfloor a_i \rfloor - \gamma r_i, \ldots, \lceil a_i \rceil + \gamma r_i\}$. For example, $\gamma = 0$ means that value a_i must be either $\lfloor a_i \rfloor$ or $\lceil a_i \rceil$, that is the zero-restricted version. Of course, the loss of information of a pattern increases when γ is larger, and for that reason it is highly recommended to use the smallest value of γ that yields a feasible pattern.

4.2.4 Cell Perturbation

The main disadvantage of the (zero-restricted) controlled rounding method is that a sufficiently disclosure-limited pattern does not always exist due to the tight constraints. Increasing γ may be undesirable because it may drastically lower the utility of an output. Therefore, a preferable way of ensuring the existence of disclosure-limited patterns is to relax conditions (4.9) and to search for a congruent table $\boldsymbol{v} = [v_i : i \in I]$ such that

$$v_i \in [\lfloor a_i \rfloor, \lceil a_i \rceil] \tag{4.10}$$

where $\lfloor a_i \rfloor$ and $\lceil a_i \rceil$ are given in advance from the DSO such that $\lfloor a_i \rfloor \leq a_i \leq \lceil a_i \rceil$. These extreme values can be defined as the nearest numbers to a_i which are multiples of a given number (as defined in controlled rounding), but they can also be the two values which are at a given distance from a_i (i.e., $\lfloor a_i \rfloor := a_i - t_i$ and $\lceil a_i \rceil := a_i + t_i$ for a given base number $t_i > 0$). Table 4.8 shows a possible pattern for Table 4.1. Table $\boldsymbol{v}$ is then a pattern in the *cell perturbation method* (also called *partial controlled rounding*, Salazar (2005)).

Table 4.8 Cell perturbation pattern for Table 4.1

	A	B	C	Total
Activity I	20	50	10	80
Activity II	7	16	26	49
Activity III	18	35	8	61
Total	45	101	44	190

As with the controlled rounding method, the loss of information of a cell i can be defined to be proportional to $|v_i - a_i|$, so the "loss of information" of a pattern measures a distance between the published table $\boldsymbol{v}$ and the original table $\boldsymbol{a}$. However, this is a poor criterion because, if all constraints (4.9) are removed and no new one is required, then the valid pattern with minimum loss of information is the nominal table, that is $\boldsymbol{v} = \boldsymbol{a}$. Hence, some constraints from (4.9) must remain (for example, the one concerning the risky cells) or, alternatively, the published values in each risky cell must be equal to some given values. An alternative procedure is to define the "loss of information" as a distance between the pattern $\boldsymbol{v}$ and a given array $\boldsymbol{a}'$, as described in the controlled rounding section. By appropriately choosing $\boldsymbol{a}'$, the objective function in the model will steer the optimization algorithm to find a valid pattern with better properties.

Let $r_i := \lceil a_i \rceil - \lfloor a_i \rfloor$ be known by the data snoopers. The data snooper problems associated with data snooper k are now exactly the same as in the controlled rounding method. A necessary (but not sufficient) condition for feasibility is that max $\{\mathrm{UL}_i^k + \mathrm{LL}_i^k : k \in K\} \leq 2r_i$ for all $i \in I$. See Salazar (2008) for technical details.

4.2.5 *All-in-One Method*

SDL may combine different methods. For example, suppose there is a partition of the cell set I into I_1 and I_2, and the DSO is interested in publishing intervals $[y_i^-, y_i^+]$ when $i \in I_1$, using interval publication, and publishing perturbed values v_i when $i \in I_2$, using cell perturbation. Then the combined method can be mathematically modeled by observing that the feasible region of the data snooper problems associated to data snooper k is

$$\begin{aligned}
\sum_{i \in I} m_{ij} y_i = b_j \quad & \text{for all} \quad j \in J, \\
v_i - r_i \leq y_i \leq v_i + r_i \quad & \text{for all} \quad i \in I_1, \\
y_i^- \leq y_i \leq y_i^+ \quad & \text{for all} \quad i \in I_2, \\
\mathrm{lb}_i^k \leq y_i \leq \mathrm{ub}_i^k \quad & \text{for all} \quad i \in I.
\end{aligned}$$

The previous four methods can be combined into the more general one because all the methods are based on the same underlying concepts, as given in Section 4.1. All the methods protect by introducing uncertainty, as required by the DSO through the given protection levels. None of the methods presented in this section require the disclosure auditing phase because the disclosure limitation, as technically defined here, is mathematically guaranteed in all feasible solutions.

4.3 Other Methods

The methods introduced in Section 4.2 always ensure that the output patterns (if any exist) are disclosure-limited in accordance with the definition given in Section 4.1. The reason for this assurance is that the requirements are built into the constraints of the mathematical programming formulation. Therefore the auditing check described in Section 4.1.8 is unnecessary on the solutions coming from these methods. However, other authors have proposed alternative approaches which have intuitive appeal, and in some cases may have better (disclosure risk, data utility) characteristics. This section summarizes the most important of these approaches, namely, table redesign, introducing noise to microdata, data swapping, cyclic perturbation, random rounding, and controlled tabular adjustment.

4.3.1 Table Redesign

Perhaps the oldest technique to protect a table having too many risky cells is *Table Redesign* (also called *Global Recoding*) which simply combines some rows or columns. Instead of eliminating risky cells with, for example, the cell suppression method (Section 4.2.1), table redesign merges different cells (for example, collecting different columns of a two-way table). As a result the combined cells aggregate data from more respondents. This can overcome confidentiality concerns about small cell counts and so yield tables deemed safe. Of course, merging information across cells does lead to information loss and this can be significant.[3] As a simple example of how a redesigned table could lose all data utility, consider a 2×2 frequency table with attributes of sex and college degree status. It cannot be redesigned by combining rows or columns without losing essentially all information about the relationship between the two attributes. Because of such concerns, table redesign should only be used after careful determination that data utility is not excessively diminished. This could be the case, for an important example, if a numerical attribute like income was banded in such narrow intervals that combining adjacent intervals would not appreciably change inferences drawn from the table. This will not be evident a priori; it needs to be empirically checked.

4.3.2 Introducing Noise to Microdata

In 1996, the US Bureau of the Census provided an alternative to cell suppression for use with establishment tabular data. With an aim of publishing more information and fulfilling more requests for special tabulations, they introduced noise to the underlying microdata, thus perturbing each respondent's data without having to suppress cells in the tabular data. Acting on the underlying microdata in this way is sometimes referred to as *pre-tabular SDL*. Evidently, injecting sufficient noise can increase a data snooper's level of uncertainty, but it begs the question of how much the data utility is reduced.

[3] Arguably, table re-design is just a special case of the table design and selection process. However, the distinction between the design and re-design processes is that re-design changes the table that the DSO would like to release, whereas the initial design phase selects that table. From a data utility point of view this is an important distinction.

A partial answer to this question is to preserve aggregate estimates that would not be risky for disclosure. Evans et al. (1998) tested this by forcing certain statistical estimates after noise injection to equal their values before noise injection. Interior table cells were then proportionally adjusted to these aggregate values. Also, as appropriate for some surveys, Evans et al. (1998) propose a hybrid technique of combining noise injection with cell suppression.

4.3.3 Data Swapping

Data Swapping is another class of pre-tabular SDL methods. A pioneer technique in this class is the *Confidentiality Edit* introduced by Griffin et al. (1989). It proceeds as follows:

1. Sample records from the microdata file.
2. Find a match for these records in some other geographic area on a specified set of important attributes.
3. Swap the geographical area codes of the matched records.

In its implementation, the US Census Bureau has increased the swapping fraction for small blocks to provide additional disclosure limitation. After the microdata file has been treated in this way it can be used without further modification to prepare tables, but still the extent of realized disclosure limitation and the impact on data utility must be empirically checked.

Data swapping was investigated by simulation in Navarro et al. (1988). It was found that swapping provides adequate disclosure limitation except in areas with small populations (say, city blocks). Data swapping is further discussed in Chapter 5 in the context of microdata.

4.3.4 Cyclic Perturbation

Cyclic perturbation is an SDL method proposed in Duncan and Roehrig (2007) for frequency tables. In the disseminated data product, the original table values are altered in a way that preserves the table's marginal totals. Further, in a Bayesian implementation of the method certain details of the procedure are made public, so that a data user can determine not only the range of values that a table cell may have had in the original table, but also the exact posterior distribution over those possible values, given the data user's prior probabilities over a relatively small set of possible original tables. This permits a valid Bayesian analysis of the published table.

Although cyclic perturbation can be extended to three-dimensional tables, we discuss it in its two-dimensional application. To protect the interior cells, cyclic perturbation modifies their values in a principled way, by applying a sequence of random perturbations to them. These perturbations leave the marginal totals unchanged. Each perturbation modifies a patterned collection of four or more cells—called a *data cycle*—with some cell values increasing by one and others decreasing by one, in such a way that each row and column sum is undisturbed. The four cells of a data cycle are alternatively signed with – and +. Table 4.9 shows three data cycles for Table 4.1. Note that each internal cell in the table appears in two cycles, which implies that each internal cell will have two chances to be perturbed. In a general situation one needs to select the cycles so each internal cells appears in the same number of cycles.

Table 4.9 Collection of data cycles for Table 4.1

	A	B	C	Total
I	+	−		80
II		+	−	49
III	−		+	61
Total	45	101	44	190

Cycle 1

	A	B	C	Total
I	−		+	80
II	+	−		49
III		+	−	61
Total	45	101	44	190

Cycle 2

	A	B	C	Total
I		+	−	80
II	−		+	49
III	+	−		61
Total	45	101	44	190

Cycle 3

Once a set of data cycles has been chosen to protect the original table, cyclic perturbation proceeds by applying each cycle once, in a fixed order, to the table. The order in which the data cycles are chosen is critical in determining which tables can be reached through cyclic perturbation. At each step, the procedure flips a three-sided coin, whose probabilities for sides A, B, and C are α, β, and $\gamma = 1-\alpha-\beta$, respectively. If the coin shows side A, cells in the positions marked + in the data cycle are increased by one, and cells in positions marked – are decreased by one. If the coin shows side B, the reverse happens. If the coin shows side C, the table is unchanged. Each such move, and so any sequence of such moves, leaves the row and column sums unchanged. The parameters α and β can be chosen sufficiently large to provide adequate disclosure limitation for the interior cells.

4.3.5 Random Rounding

A technique similar to controlled rounding is random rounding, where firstly the internal cell values of the table are rounded up or down, and secondly the marginal cell values are recalculated using the modified internal values. A clear advantage of random rounding is that it is simple. Another advantage is that we can decide for each internal cell whether to round up or down with a certain probability such that the result is an unbiased table. The clear disadvantage is that a marginal cell can have a published value far from its original value, thus the utility of the publication may be too small. This disadvantage is not present in controlled rounding.

Random rounding is a perturbation technique that rounds cell values up or down with a certain probability. It consists of replacing each value a_i in the original table by the result of the following function**:**

$$v_i := \begin{cases} \lceil a_i \rceil & \text{with probability}\, p_i := (a_i - \lfloor a_i \rfloor)/r_i; \\ \lfloor a_i \rfloor & \text{otherwise;} \end{cases}$$

where $\lfloor x \rfloor$ is the largest integer smaller than x and $\lceil x \rceil$ is the smallest integer greater or equal than x, as defined in Section 4.2.3. An inconvenience of this simple procedure is that the output table $[v_i : i \in I]$ is likely to be a non-congruent table, that is, it does not satisfy condition (4.1). This feature can be perceived negatively by data users. A table prepared using random rounding could lead an uninformed public to lose confidence in the numbers: at worst it looks as if the DSO cannot add! Another problem is that a rounded table from this procedure may not be disclosure limited, by the definition given in Section 4.1. Therefore a disclosure audit (see Section 4.1.8) is required.

Apart from its simplicity, the main advantage of random rounding is a statistical property of the output: by construction, a rounded table is unbiased. This is not the case with *deterministic rounding*, where the procedure replaces the value a_i by

$$v_i := \begin{cases} \lceil a_i \rceil & \text{if } 2a_i \geq \lceil a_i \rceil + \lfloor a_i \rfloor \, ; \\ \lfloor a_i \rfloor & \text{otherwise.} \end{cases}$$

With deterministic rounding, bias is produced by the fact that small values are rounded down, and big values are rounded up. We do not recommend this procedure.

Another idea is to randomly round the internal cells of a table, and then compute the marginal cell values directly from the rounded internal cell values. The disadvantage with this is that values of marginal cells may be at some distance from their original values. A way to avoid this drawback is investigated in Salazar (2006b). Each internal cell i is associated with a decision variable 0-1 ILP problem is then solved to reduce the distance between the rounded and the unrounded value for all the marginal cells. Salazar (2006b) shows that acceptable rounded tables can be found on tables with up to 1 million cells and complex structures in reasonable time.

4.3.6 Controlled Tabular Adjustment

Dandekar and Cox introduced a perturbation method for tabular data known as *Controlled Tabular Adjustment* (CTA); see, for example, Cox et al. (2004). Risky cell values are replaced by either of their closest safe values and adjustments, as small as possible, are made to other cells to restore the table additivity, so that marginal totals are maintained. As with other methods we have discussed, mathematical programming techniques are used to implement CTA. The technical formulation follows:

For notational convenience, let $\mathrm{ll}_p^1 := a_p - \mathrm{LL}_p^1$ or $\mathrm{ul}_p^1 := a_p + \mathrm{UL}_p^1$. Another assumption is that the data snooper will not compute intervals of potential values for each risky cell. Then the CTA method replaces each risky value a_p by either $\mathrm{ll}_p^{\mathit{1}}$ or $\mathrm{ul}_p^{\mathit{1}}$. To restore additivity and produce a congruent table, some non-risky values may also need to be replaced. The problem is now to find the new values in such a way that the original and modified tables are as close as possible. Since no other requirement is considered, the CTA problem of finding the new values can be modeled by the mixed integer linear program illustrated in Fig. 4.1. In this model, the

$$\sum_{i \in I} w_i^+ z_i^+ + w_i^- z_i^-$$

subject to:

$$\begin{array}{rll}
\sum_{i \in I} m_{ij}(a_i + z_i^+ - z_i^-) = b_j & & \text{for all } j \in J \\
z_i^+ \geq UL_i^1 x_i & & \text{for all } i \in P \\
z_i^- \geq LL_i^1 (1 - x_i) & & \text{for all } i \in P \\
x_i \in \{0,1\} & & \text{for all } i \in P \\
0 \leq z_i^+ \leq ub_i^1 - a_i & & \text{for all } i \in I \setminus P \\
0 \leq z_i^- \leq a_i - lb_i^1 & & \text{for all } i \in I \setminus P.
\end{array}$$

Fig. 4.1 Mixed ILP model for controlled tabular adjustment

decision variable x_i assumes value 1 if the risky cell i should be increased to $\mathrm{ul}_i{}^{1}$, and 0 if it should be decreased to $\mathrm{ll}_i{}^{1}$. The variables $z_i{}^{+}$ and $z_i{}^{-}$ represent the positive and negative adjustments, respectively, that need to be applied to a cell i before publication.[4]

Comparing CTA with the cell perturbation method, we note that both methods search for a congruent table as close as possible to the original one. But in addition, cell perturbation guarantees all the disclosure limitation level requirements as defined in Section 4.1. This consideration justifies the large number of linear inequalities in the mathematical model associated with the DSO problem in cell perturbation. These constraints are not considered in the CTA model and therefore, if the CTA problem has an optimal solution with a_p replaced by $\mathrm{ul}_p{}^{1}$ (i.e., with $x_p = 1$), then it is not guaranteed that by publishing this solution the data snooper will find another congruent table with a value at most $\mathrm{LL}^{1}{}_p$ in cell p. Hence, CTA only guarantees either the upper *or* the lower disclosure limitation level for each risky cell, whereas cell perturbation guarantees all of the upper *and* the lower protection level requirements. Furthermore, CTA looks for *one* table $[y_i : i \in I]$ satisfying *one* disclosure limitation level *for each* risky cell. Cell Perturbation searches for a table $[v_i : i \in I]$ such that, when it will be published, for each risky cell there will exist congruent tables $[y_i : i \in I]$ satisfying each disclosure limitation level requirement. As CTA searches for a table holding exactly one disclosure limitation level for each risky cell, it is likely that no values y_i can be found for all the non-risky cells such that the final table satisfying *all* these conditions *simultaneously*. In other words, the mathematical model in Fig. 4.1 is likely to be *infeasible* for complex table structures. Summarizing, on the one hand, CTA is a more relaxed method than cell perturbation because it does not require both upper *and* lower protection but only upper *or* lower protection, for each risky cell. On the other hand, CTA is a more constrained method than cell perturbation because it aims at finding a single table that must guarantee the protection of all the risky cells.

[4] As with cell perturbation, the reason for using two continuous variables instead of a single one is the technical one of enabling a linear objective function.

These observations, however, do not mean that CTA is an incorrect SDL method, but we only point out that CTA will not definitely produce optimal (or near-optimal) solutions that meet the assumptions given in Section 4.1. CTA protects a table in a more aggressive scenario where, for each risky cell p, a data snooper does not compute the disclosure-limited interval $\left[\underline{y_p^k},\ \overline{y_p^k}\right]$ from the published table and a priori knowledge.

4.4 Summary

Protecting a source table means being able to produce a publishable table from which a data snooper may only claim that there is a wide set of possibilities for the value of each risky cell. This set cannot be too small or disclosure risk will be too high. Also, it cannot be too large or loss of utility will be too high. Whilst this implies that the main SDL problem is a bi-criteria optimization problem, it is well accepted that the most convenient way to avoid the bi-criteria issue is to optimize one criteria—utility—subject to a strict limitation of the other criteria—disclosure limitation. This chapter has presented four methods to protect tabular data using this framework: cell suppression, controlled rounding, interval publication, and cell perturbation. Other methods from the literature have also been described, notably cyclic perturbation and controlled tabular adjustment.

Chapter 5
Providing and Protecting Microdata

A microdata file is a compilation of data records. Each record contains values of attributes about a single unit—say a person with the attributes of their height, attitude toward minimum wage laws, cell-phone usage, and diastolic blood pressure. Microdata are special. *Expanding Access to Research Data: Reconciling Risks and Opportunities* (National Research Council, 2005) recognizes both the enthusiasm for the research potential of microdata and the trepidation about the risk microdata pose to confidentiality. In this chapter we help reconcile the inevitable tension of both protecting and providing microdata. We lay out the principles of microdata confidentiality and identify ways of improving DSO practice.

Among DSOs, national statistical offices have led the way in developing and implementing new practices for release of confidential microdata. Statistics Netherlands (2007) provides a comprehensive treatment of confidentiality issues for microdata and Jabine (1993b) reviews the then current practices of US federal statistical agencies. McMillen (2001) describes practices of a particular agency. United Nations (2007) with electronic updates[1] provides a comprehensive view of current confidentiality practice regarding microdata, especially from the perspective of European National Statistical Offices. It includes case studies for a number of individual countries that discuss what they consider to be good practice, who their target audience is, and if there is any supporting legislation. Specifically, cases are provided about the following topics: legislation to support release of microdata for Australia and Finland; data cube (high-dimensional tabular data) for the Netherlands; public-use microdata for the United States; release of licensed microdata files for Australia, the Netherlands, and Sweden; remote data access for Canada, Australia, and Denmark; research data centers for Canada; data laboratory microdata access for the Netherlands, New Zealand, Brazil, and Italy; and managing decision making on confidentiality for Slovenia and Australia.

Confidentiality researchers have developed a substantial literature about protecting microdata. Within this literature the following works deal with general issues: Duncan and Lambert (1986, 1989) provide a decision-theoretic foundation for

[1] www.unece.org/stats/publ.htm

G.T. Duncan et al., *Statistical Confidentiality*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-7802-8_5,

microdata protection. In the same spirit is Zaslavsky and Horton (1998). Doyle et al. (2001) provide a broad compendium of SDL methods, including ones for microdata, as does Federal Committee on Statistical Methodology (2005).

In Chapter 1 we established how DSOs provide data that serve the information needs of societies, especially those with the particular demands of democratic and free-market institutions. Further, we argued that these benefits could not be achieved without protecting confidentiality. In Chapter 4 we described techniques for protecting tabular data. Conceptually, these methods are similar to those that can be used to protect microdata. They include perturbation, coarsening, and combining. They also include the generation of synthetic data. But in this chapter we ask a new question: What is of special concern for microdata?

Microdata release heightens both risks and rewards. Consider an actual survey conducted in the aftermath of Hurricane Georges that struck the Dominican Republic in 1998. The storm took 300 lives, damaged infrastructure, and destroyed crops. To assess relief needs, the US Centers for Disease Control and Prevention (CDC) and the American Red Cross surveyed 33,000 households. The resulting microdata included such attributes as the number of days per week that the household had insufficient food (both before the hurricane and after the hurricane) and whether the household had someone with a gastrointestinal illness. Besides providing the facts needed to mount a relief effort in the aftermath of Hurricane Georges, such data help analysts answer research questions such as “What are the most immediate needs of households in the aftermath of a hurricane?” and “Could stockpiling critical supplies alleviate suffering?” Disaster relief organizations need the answers to such questions so that they can properly plan and manage their operations.

DSOs broker the provision of microdata from data providers to data users. The microdata may come from a survey (like the CDC survey in the Dominican Republic), a census (like the Population and Housing Census in Spain), or from administrative activities (as with issuing driver’s licenses by the State of New Mexico’s Department of Motor Vehicles). Microdata are essential to analysts, planners, and researchers. Thus microdata have high data utility, and with inexpensive electronic record storage may indeed retain that value into the future, as the 1998 CDC survey had value for relief organizations in responding to the needs of people affected in 2005 by the far more devastating Hurricanes Katrina and Rita that struck the Gulf Coast of the United States.

Given the value of microdata, why should not the CDC just move such microdata as expeditiously as possible to a broad range of researchers? As the reader surely knows at this point, the obstacle is confidentiality. After laying out issues and concerns in Sections 5.1, 5.2, and 5.3, this chapter shows in Sections 5.4 and 5.5 how the confidentiality obstacle can be hurdled. Each section of this chapter addresses a different aspect of the utility of microdata and means of confidentiality protection:

- *Section 5.1* confirms the contribution of microdata and clarifies what users need in such data. Given this justified demand for microdata, a DSO cannot just refuse to provide it and still satisfy its mandate to provide useful data.

- *Section 5.2* identifies the ethical, pragmatic, and legal considerations that motivate a DSO's confidentiality promises to data providers. As we argued in Chapter 1, by ignoring these promises, a DSO puts its mission at risk. To appreciate where disclosure risk arises for microdata, threats to confidentiality are identified.
- *Section 5.3* points to the characteristics of microdata that make them vulnerable to confidentiality attack.
- *Sections 5.4–5.12* explore various masking methods. This gives us a range of procedures for SDL that can be effective in protecting confidentiality.
- *Section 5.13* introduces synthetic data, that is, data that are stochastically generated from a model inferred from the source data. This gives us another way of providing users with confidentiality-protected microdata.
- *Section 5.14* concludes this chapter with thoughts about the state of the art in providing microdata under confidentiality constraints.

5.1 Why Provide Access?

The question of why DSOs should provide access to microdata has an obvious answer—something like, "today's researchers require it." Certainly this answer is a true statement about researchers' needs. To give one example, a report by the UN Statistical Commission and the Statistical Office of the European Communities states that "Statistics Netherlands has seen an increasing demand for dedicated analyses [of microdata] to be performed from policy-makers and from the research community."[2] If the simple fact of researcher demand were a sufficient answer, there would be no need to read this section. But this answer is incomplete, and certainly insufficient for a DSO trying to create coherent policy about data access. After all, as Joris Nobel notes for Statistics Netherlands (but certainly applicable to most DSOs), "There is no general obligation to transmit microdata."[3]

A short, and much better, answer to why DSOs should provide access to microdata is that it serves important societal needs for evidence-based analysis, and these needs cannot be served in other ways. See, for example, Spruill (1983) and National Research Council (2005). In this section we explore the distinctive value of microdata for research and policy analysis.

Providing access to microdata has several advantages over just providing statistical aggregates:

1. More so than with other data products, microdata allow researchers and analysts to study complex issues based on factual evidence. Hence microdata make contributions to public debate, research, and decision making.

[2] www.unece.org/stats/documents/ece/ces/sem.54/13.e.pdf

[3] unstats.un.org/unsd/goodprac/bpform.asp?DocId=244&KeyId=20

2. Microdata permit new data analyses and data mining. Thus microdata enhance the potential for data integration that can link areas that may have only previously been studied separately, for example maternal health and juvenile delinquency.
3. Microdata increase user trust in the validity of aggregate statistics.
4. Microdata enhance public recognition of the value of official statistics and so develop a core constituency for the DSO.
5. Having actual microdata values helps researchers provide feedback to DSOs that can contribute to improvements in data quality.

The set of those who require access to microdata certainly includes researchers in universities, but it also includes policy analysts working in non-government organizations and in government agencies. For many purposes of this broad data-using community, microdata cannot be replaced with aggregate data. In Chapter 4, we noted that many of the products of statistical agencies are tables and various forms of summary statistics. As useful as tabular data have proven to be, microdata are needed for analysis because of the following:

1. Many important questions in fields such as demography, economics, and epidemiology can only be answered based on microdata. Foster et al. (2001), for example, found from analysis of microdata that the widespread reallocation of factors of production from one firm to another firm is a major contributor to US productivity growth—indeed more important than investment in equipment and structures. This result could not be obtained from macro-level indicators.
2. Microdata can provide the information about geography of small areas and detailed timing that are essential to understand phenomena such as job mobility and access to emergency medical treatment. Longitudinal microdata can provide information about duration in states (such as "unemployed" or "in ambulance to hospital" and transition between states (see, for example, Steel and Sperling (2001)).
3. Hierarchical microdata (for example, people in households, employees in firms, or children in schools) allow multilevel analyses, which take account of variability at each level. As Goldstein (1987) demonstrates, without such data incorrect inferences can be made.
4. Statistical models require microdata for direct inference and validation. Multiple regression models, specifically, have substantive and policy relevance since they allow the estimation of the effects of changes. As emphasized by Lane (2003),[4] the work of Lee et al. (1999) using microdata showed that the marginal effect of income on Medicare expenditures is broadly positive for men, but that the relationship is much flatter for women.
5. Many alternative statistical models cannot be explored and compared without microdata. For example, inference about nonlinear regression models requires microdata.

[4] www.unece.org/stats/documents/ces/2003/crp.2.e.pdf

6. Researchers can analyze subgroups of the population based on their own selection from the individual records.
7. Data quality cannot be independently assessed without access to microdata. Replication and comparison with other data sources encourages accountability of data stewardship organizations. This provides a scientific safeguard, a counterbalance to political pressures to obtain specified results. It also encourages improvements in survey and census methodology.
8. Microdata allow linkage to other microdata that can enrich analysis possibilities.

As a specific illustration of the importance of microdata, Lane (2003) noted:

> One common characteristic of the statistical institutes of the countries in which I worked was a reluctance to provide access to microdata—and in every case, this led to incomplete analysis and wasted resources in countries that could afford them least. In one case, the country in question was concerned about the low labor force participation rate of women—which had hampered development for over a decade. Several policy options were on the table—including providing free childcare, flexible work-weeks, and subsidized education. However, no micro-data analysis had been undertaken. Although labor force surveys were regularly fielded, they were not even released to the Ministry of Human Resources or the Ministry of Education. We analyzed the micro-data and found that, even after controlling for education, industry and occupation, women were paid 60% less than men—and had been for the ten years in question. Our conclusion, which would have been apparent to any analyst working with these data, was that the country in question would have been better served by investigating the sources of these earnings differentials, rather than investing in the expensive set of options initially identified. Had the country in question permitted broader access to the micro-data a decade earlier, the appropriate policies could have been in place much earlier.[5]

National statistical offices have begun to release public-use microdata files. Historically, the first microdata files were released from the 1960 US Census (with retrospective microdata files later extracted for earlier years). The microdata were first called the Public Use Sample (PUS), and (happily) renamed the Public Use Microdata Sample (PUMS) in 1980. Canada first released public-use microdata files (PUMFs) from the 1971 Census and has continued this policy for every quinquennial census since. Australia first produced microdata files for its 1981 Census. In collaboration with Member States and participating countries, Eurostat releases microdata products from several surveys including the Community Innovation Survey, the European Community Household Panel, and the Labour Force Survey.

In the United Kingdom, where the practice of releasing samples of anonymized records was accepted in 1989, the Office of National Statistics has noted that microdata have been the basis for research into diabetes, infant mortality, the health of the nation, demographic changes, the labor market, identification of trends, and government planning.

Echoing this affirmation of value, Statistics New Zealand—the national statistical office—states,

[5] www.unece.org/stats/documents/ces/2003/crp.2.e.pdf

> As part of its mission to 'provide New Zealanders with a national statistical service of integrity that is valued and trusted for publishing useful information', Statistics New Zealand wants the best possible use made of the data it collects. Microdata access is an increasingly important input to social science and economic research.[6]

Statistics Canada provides the Census Public Use Microdata Files (PUMF),[7] because of the following:

1. They provide access to a comprehensive social and economic database about Canadians.
2. They are a research tool unique among their census products in that they give users access to non-aggregated data. The files contain attributes that data users can manipulate according to their own requirements.

Table 5.1 gives several applications—their nature, which DSOs have responsibility, what data providers are the source of supply, what data users are the source of demand, and to what use the microdata are put. This table gives a sense of why microdata are needed to support policy and management analysis.

Table 5.1 Examples of supply and demand for microdata

Data stewardship organization	Supply: the data providers	Microdata	Demand: some data users	A use
US Social Security Administration	Claimant	Master beneficiary record—insurance amount, name, date of birth, sex, date of death, date of filing, relationship to the SSN holder, benefit amount, and payment status	Economists and policy analysts	Economic and demographic analysis of the future solvency of the social security system
Office of National Statistics, United Kingdom	UK firms	Inter-departmental business register	Salisbury District Council	Land use planning and economic development functions
Office of National Statistics, United Kingdom	UK residents	Census	University of Manchester	Labor market circumstances of minority ethnic communities
US National Center for Health Statistics	Respondents to National Health Interview Survey	Demographics, health status	Epidemiologists	Americans' knowledge of cancer risk

[6] www.stats.co.nz/about_us/policies-and-guidelines/microdata-access-protocols.aspx

[7] www.statcan.gc.ca/bsolc/olc-cel/olc-cel?lang=eng&catno=95M0028X

For example, we see that the Salisbury District Council uses data products provided by the United Kingdom's Office of National Statistics. With these survey data on intentions of firms, indications of expansion of employment activity can be integrated with environmental factors, like the flood plain of the River Avon, and aesthetic factors—say good views of the Cathedral—to produce a land use plan.

So, from the perspective of data users, there is much to be gained from access to microdata (Dale, 1998). With it they can develop statistical models that incorporate the complexity of social and economic interactions, they can test these models against empirical evidence, they can identify policy levers, and they can make better predictions. Given the substantial expenditures that government agencies and private organizations make in data gathering, full use of microdata recoups that investment.

As appealing as these benefits are and no matter how much a DSO is motivated to facilitate them by moving microdata to the user, the threat to confidentiality remains. In Section 5.2 we identify reasons for protecting this confidentiality.

5.2 Confidentiality Concerns

As is evident from the theme of this book, data stewardship requires attention to confidentiality. Difficult as this may be for aggregate data, such as the tables examined in Chapter 4, microdata have characteristics that lead to even more disclosure risk. Literature relevant to the disclosure risk of microdata is voluminous. Sources worth examining include Béland (1999), Blien et al. (1992), Carter et al. (1991), De Waal and Willenborg (1996), Mokken et al. (1992), and Willenborg and de Waal (1996, 2001). Also see the material[8] on the Eurostat website.

Protecting microdata is impossible if the record includes *direct identifiers*—like social security numbers or establishment id numbers. Maintaining confidentiality is also problematical with *close identifiers*, such as names, physical addresses, and IP addresses. Direct identifiers are mapped on a one-to-one basis with the data subject, and so no additional information is required to disclose their identity. Once the identity of the record is disclosed, all the information on the record is disclosed, whether it is deemed sensitive or not. In many surveys, information about other household members (ages, household income, housing characteristics) would be revealed as well because they are part of the data record.

Therefore, a necessary starting point for protecting data is striking all direct identifiers from the records before they are released. Specifically for personal health information, HIPAA regulations under The Privacy Rule of 2003 require the removal of 18 different types of identifiers. These include obvious ones, such as names and social security numbers, but also less obvious ones, such as birth date, vehicle serial numbers, URLs, and voice prints. But such deidentification may not be sufficient to prevent reidentification by a determined and resourceful data snooper (see Elliot and Dale (1999) and Fienberg et al. (1997)).

[8] epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/statistical_confidentiality/confidential_data/introduction

As noted in Chapter 2, even after *deidentification*—the removal of direct and close identifiers—microdata can be easy targets for a data snooper. The risk of disclosure can be high because the data snooper may be able to *reidentify* some records. One reason for this vulnerability is public availability of identified data and rapid advances in the technology of linking files (Winkler, 1993, 1998; Gomatam and Larsen, 2004). File linking is possible when identified data sets are available to the snooper and these data sets share with the released file both individual respondents and certain attribute variables. We call such a data set an *identification file*. Record linkage (Fellegi and Sunter, 1969; Winkler, 1994, 1995a, 1998, 2004a, b) may be possible between the released file and the identification file using the shared attribute variables as a key.

As illustrations of how reidentification through record linkage can come about, consider the following examples of what could be identification files:

1. InfoUSA can provide extensive mailing lists according to a user's specification. A user specifies, say, a zip code (87501), an age range (50–70), and a value of house owned (over $300,000). InfoUSA provides a mailing list with each individual's credit information, marriage date, ethnicity, date of birth, and mortgage information. Such a mailing list has many key variables, especially address, ethnicity, and data of birth that allow linkage to many social surveys.
2. Dun and Bradstreet can provide the identities of the top 25 florists in the United States, together with sales figures, number of employees, and credit information. The availability of such rich information makes the presence of a unique combination of attributes and a linkage to establishment survey data more likely.
3. In the United States, the Social Security Death Index contains variables of sex, inferred age, and death date. An MIT student paper[9] used these variables as key in linking to the deidentified Chicago Homicide Database. It showed that a high percentage of victims could be reidentified.

Attention, therefore, must be paid to protecting microdata through SDL methods. Describing how such methods work with microdata is one of our goals in this chapter and will be addressed in Sections 5.4 and 5.5. When SDL has been successful in reducing disclosure risk sufficiently so that reidentification is essentially impossible, the resulting data product is said to be *anonymized*.

Here are some fundamental definitions that will provide us with a common language:

Identification. Linking of a data record to a particular population unit.

Deidentification. Removal of direct identifiers, such as name, Social Security Number, or e-mail address in a data record.

Reidentification. Linking a direct identifier to a deidentified data record.

[9] http://web.mit.edu/sem083/www/assignments/reidentification.html

An *identification file* is a database of microdata records that:

1. is available to the data snooper.
2. has structural identifying information, such as name, address, or social security number.
3. has attribute values that overlap with the attributes for the data snooper's target records in the source data that the data stewardship organization wants to protect.

Confidentiality is an assertion that sensitive objects will be protected. One class of sensitive objects is made up of attribute values, for example, an individual's net worth or HIV status. Determining which attribute values are to be considered sensitive is a non-trivial task, since sensitivity is context-dependent. For example, marital status is in many cases public knowledge, whilst in certain welfare cases or in applying for a marriage license it may be pivotal to qualification and therefore sensitive.

As we have noted, deidentification of a record is insufficient to prevent reidentification. There are at least seven features of microdata that, because they increase disclosure risk, challenge the assertion of confidentiality: Some of these features are inherent to the data release situation (linker files), others are commonly encountered for microdata files (many attributes), and some may or may not be a feature of a particular file (longitudinal data).

Existence of identity linker files. In this electronic era, a DSO should anticipate that a data snooper will have ready access to databases of identified records, rich in number of records and attributes. Common data sets that include identifiers are *marketing databases*. Such databases also include other attribute variables about customers and prospects, some of which may overlap with those in the subject database. Today, commercially available record-linkage software can make it easy to match records across databases.[10] Once a link has been established (and verified), an identity disclosure has occurred.

Geographical detail. Typically, researchers want data with as much geographical detail as possible. However, a record that includes an attribute specifying that the data subject is associated with a small area, say lives there or works there, will tend to increase disclosure risk. In response to this specific confidentiality concern, DSOs usually only disseminate data with coarse geographical codes. For example, the smallest region for which the US Census Bureau tabulates 100% data is the *census block*, which in urban areas may consist of a single city block or in rural areas may cover several square miles. The Census Bureau imposes minimum size constraints because in an area with so few inhabitants, several records may be unique. Then the probability of being able to match a single record against an external data source having identifiers may be high. See, for example, Greenburg and Voshell (1990), Elliot et al. (1998), and Steel and Sperling (2001) for detailed discussion of the relationship between disclosure risk and geographical detail.

[10] http://www.fcsm.gov/working-papers/charlesday.pdf

Longitudinal or panel structure. Increasingly researchers are requesting longitudinal data for their analyses. Longitudinal data sets have measurements for single study units over multiple periods of time. With data collected on a single observational unit over time, there may well be few records with the same combinations of attribute values. For example, many electrical engineers lived in the Rome area in 1998 and many electrical engineers lived in the Milan area in 1999, but few did both. So an electrical engineer who has moved from Milan to Rome in 1999 will be rare within a data set which traces people over time. Because of this particularity, longitudinal data pose substantial disclosure risk. The Disclosure Review Board for the U.S. National Center for Health Statistics warns that when a related identity linker file exists that the potential for matching increases "... if longitudinal data are being collected (i.e., if the data for the same respondents/units will be collected for several different reference periods). Primary concern relates to time series of data items potentially linkable to outside records (e.g., income tax or employment records)." Trottini (2005) argues for a decision-theoretic formulation of confidentiality problems involving longitudinal data. As a specific but important example of longitudinal data, *panel data*, also called cross-sectional time series data, are data where multiple cases (people, firms, countries, etc.) are observed at two or more time periods. An example is the U.S. National Longitudinal Survey of Youth, where a nationally representative sample of young people is surveyed repeatedly over multiple years.

Outliers. The presence of multiple attributes gives rise to outliers even where the individual values might be unexceptional and these can lead to identifiable data subjects. For example, Fig. 5.1 plots the log of concentration of plasma triglycerides against plasma cholesterol for 320 patients having narrowing arteries. It shows why outliers require protection. A data snooper might know that a patient in these data had especially high levels of triglycerides and relatively low levels of cholesterol.

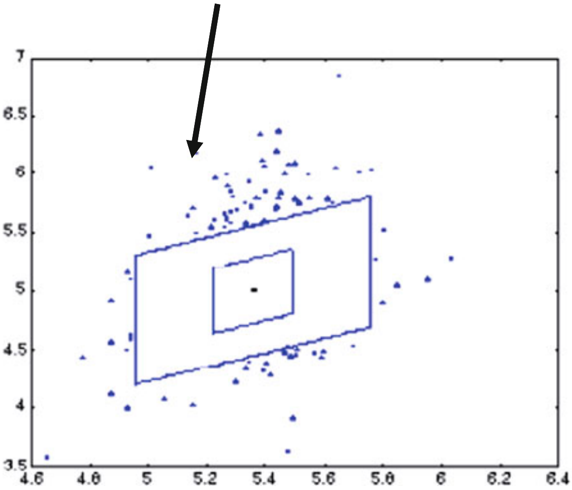


Fig. 5.1 Outliers matter

With the release of these data, the data snooper gains knowledge of the exact values of both. As just one example from the literature, Mateo-Sanz et al. (2004) deals with outliers from a confidentiality perspective.

Many attributes. If a record is rich in attributes, the chances increase that some can be matched to an identity linker database.

Detailed attribute variables. Attribute variables are detailed if values are finely coded, for example, income coded to the nearest dollar or religious affiliation disaggregated to the denomination level—say, Sunni Islam, Disciples of Christ, or Conservative Judaism. This detail increases the chance of linking. Further, with fine resolution on attribute variables, entities that are unique in the sample are more likely to be unique in the population (Zayatz, 1991).

Census versus survey/sample. With release of a population census, a record that links uniquely is surely identified. With survey data or the release of only a sample of data, a linked record that is unique in the sample, may well not be unique in the population, and so is not surely identified. Lack of population uniqueness is more likely if the data came from a survey with a small sampling fraction. In the United Kingdom, the Office of National Statistics (ONS) makes use of Sample of Anonymised Records (SAR) to lower disclosure risk from census microdata. Based on an analysis of disclosure risk by Dale and Elliot (2001), ONS proposed to raise the sample size of the SAR from a 2% level to a 3% level.

5.3 Why Protect Microdata?

In Chapter 1 we argued that statistical data must have confidentiality protection because of ethical, pragmatic, and legal considerations. From the previous section we know how microdata are vulnerable to confidentiality attack. But should the DSO seek to protect the microdata? As with any security issue, we protect when the object is under threat, subject to harm, and capable of being protected without losing value. For microdata, we argue that these three conditions are fulfilled:

Under threat. Many DSOs—certainly many national statistical offices—believe that when it comes to microdata the data snoopers are not hypothetical but real; that there are individuals and organizations that would, if they could, seek to identify records and so breach confidentiality (Ruggles, 2000). According to the Privacy Rights Clearinghouse,[11] identified records are indeed sought. In the United States, for example, Department of Motor Vehicle records have been consulted by employers, insurance companies, attorneys, and private investigators. Formerly such records could routinely be purchased by marketers, but that option has been restricted. While it may seem strange to call them data snoopers, parents in the United States have the right to inspect records a school has about their child under the age

[11] www.privacyrights.org

of 18. Presumably because it would be sought, criminal history information compiled by local and state criminal justice departments is not public in California. Noting especially the constraints on employer access, "Rap" sheets (records of arrests and prosecutions) can only be obtained by:

1. Law enforcement agencies
2. Attorneys working on a case involving the individual
3. The subject of the information
4. Probation or parole officers
5. A state agency which needs the information to license an individual
6. Employers, under limited circumstances authorized by law

Subject to harm. Without doubt, statistical agencies perceive that if a breach of confidentiality were to become public, their reputation would be damaged and their prospects for collecting quality data in the future would be severely hampered. Indeed in some cases because of legal stipulations, agency staff might even be subject to criminal penalties.

Capable of being protected, without losing value. There is no point in protecting something if the very act of protection causes it to lose too much value. If the point of having a microscope is that it be used in a school science class, protecting it by always keeping it locked in the principal's office is futile. So with microdata, masking the data through disclosure-limiting transformations may fuzz the actual record values enough that these restricted data are safe and cannot be linked to identifier databases. But, in this instance, are these masked data no longer useful for statistical purposes? Alternatively, DSOs may set up administrative procedures that restrict access, for example, supervised research data centers that provide safe settings. But, in this case, are the conditions so onerous that legitimate data users are deterred from access? In both the cases, the answer may be yes, that substantial masking renders the data useless and physical and administrative constraints are indeed too troublesome. In both cases, legitimate data users are denied access to useful data. However, as we will argue in the rest of the chapter, judicious use of restricted data and restricted access can protect microdata without losing its statistical value.

Given confidentiality concerns and the ability of DSOs to protect microdata, it is not surprising that DSOs have policies for microdata dissemination. Especially with National Statistical Offices (NSOs), these policies are often specified in legislation. As noted in McCaa and Ruggles (2002), there is a great deal of variation in legislative regimes and dissemination policy. Many NSOs, including those in Argentina, Croatia, Canada, India, Spain, the United Kingdom, and the United States, do not permit release of statistical products from which a particular respondent can be identified. Some countries, for example, the Netherlands and Singapore have allowed access to identifiable data provided the respondent has given consent. Other countries, such as Norway have had explicit provision for researcher access to microdata.

Historically, confidentiality has absorbed the attention of statistical agencies at the national level. Now, globalization in general, European integration in particular, has added international considerations. Eurostat is the umbrella organization that provides statistical information across the European Union and links to national statistical offices worldwide. Research collaboration and policy studies by international organizations across national borders are now commonplace, facilitated by the Internet. Lack of common data access rules across countries, however, inhibits such collaboration and make it difficult to do cross-country comparative studies. IPUMS [Integrated Public Use Microdata Series] International[12] is an organization based at the University of Minnesota that seeks to harmonize these differences and so provide access from around the world. At the same time, IPUMS seeks to protect confidentiality according to the standards of the collaborating countries.

We emphasize that disclosure limitation methods lower disclosure risk at some cost in data utility. In implementing any such method DSOs must not only assure themselves that the released data product will protect confidentiality but also that the product will be worthwhile to data users. Thus the unnecessary or overly stringent application of disclosure limitation must be avoided. Also, DSOs should provide information for users about the inferential accuracy of their disclosure-limited data products. They should answer the question: how good is the product for drawing statistical inferences? As noted in the Royal Statistical Society's Response to the National Statistics Protocol on Data Access and Confidentiality[13]:

> It is sometimes argued that the provision of such information damages confidentiality protection. We believe that this argument has no sound general basis in the methodology of SDC [Statistical Disclosure Control], although it may be justified in special cases for pragmatic reasons. The purpose of SDC is to mask details of individual respondents not to mask the statistical characteristics of interest. There is therefore no fundamental conflict between reporting accuracy of the estimates of these characteristics and protecting the confidentiality of respondents.

5.4 Restricted Data

From Sections 5.1 and 5.2 you can see the DSO's dilemma: the source microdata are indeed valuable, but the disclosure risk in their release is too high. To provide both useful and safe data and so resolve this dilemma, according to Chapter 2 the DSO has two alternatives:

1. *Restricted access.* The DSO provides access only to those who have been vetted and only under restrictions on their access. By these administrative controls the DSO seeks to ensure *safe settings* for access to data. Such controls can include licensing, bonding, providing access only in special venues, requiring

[12] https://international.ipums.org/international/

[13] http://www.rss.org.uk/PDF/DataAC.pdf

users to have sworn agent status, and having the authority to impose legal or other sanctions on those who may abuse the privilege of access. We expand on the workings of restricted access in Chapter 7.

2. *Restricted data.* The DSO employs SDL to transform the risky source data into a data product with adequately low disclosure risk. Restricted microdata (also called *safe data*) are generally public-use files, which require no authorization and are widely disseminated to the public at large.[14]

Schematically, Fig. 5.2 contrasts restricted access and restricted data. Through its server, the DSO either supplies restricted data (safe data) to the user or responds to user requests under restricted access stipulations (safe settings).

Early work on confidentiality protection by restricted data is by Fellegi (1972). Jabine (1993a) reviews methods for producing restricted data. Also see Fienberg (1994, 1997). Direct transformations of data for confidentiality purposes are called *disclosure limiting masks*.

In the process of SDL diagrammed in Fig. 5.3, the DSO initially does an audit of the source data, both to evaluate the utility of the data and to assess disclosure risk. A *confidentiality audit* will include identification of (1) sensitive and at risk objects, and (2) characteristics of the data that make it susceptible to attack. Given appropriate design and implementation of data capture, data utility is high. On the other hand, release of source microdata typically results in excessively high disclosure risk. This vexing problem can be expected in either of these two situations:

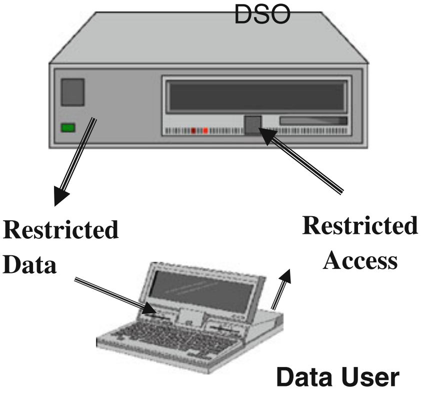


Fig. 5.2 Confidentiality protection through safe data (restricted data) and safe settings (restricted access)

[14] Some DSOs release a wider range of data products by combining aspects of restricted data and restricted access. Examples of this are the special license SARs in the United Kingdom and the Remote Access Data Laboratory (RADL) which is a secure online data query service that approved clients may access via the Australian Bureau of Statistics website to the Confidentialised Unit Record Files (CURFs).

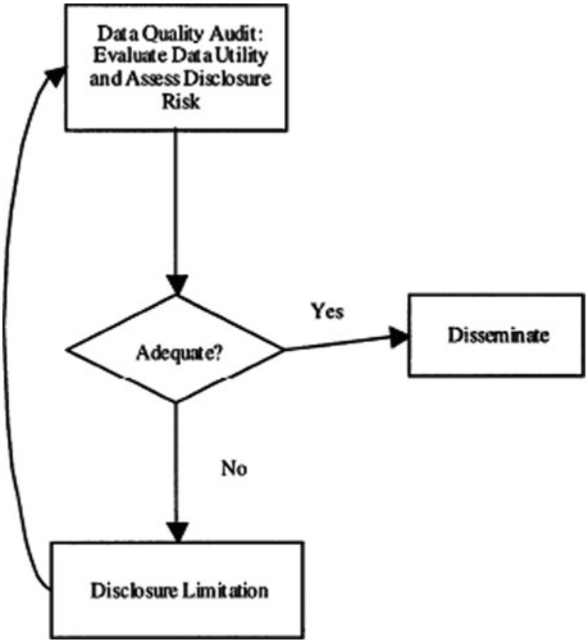

Data Quality Audit: Evaluate Data Utility and Assess Disclosure Risk
Adequate?
Yes
Disseminate
No
Disclosure Limitation

Fig. 5.3 SDL process

1. *There are univariate identifiers on the record.* A confidentiality problem obviously occurs when certain attributes give identifying information for the subject of a record. This would certainly be the case with names and addresses, whether physical, e-mail, or IP address, but it might also be the case with automobile license plate numbers, biometric identifiers such as fingerprints and facial photographs. The U.S. HIPAA [Health Insurance Portability and Accountability Act] Privacy Rule specifies 18 classes of identifiers that include the above, but also many others such as medical record number, health plan beneficiary, all geographic subdivisions less than a state, and all elements of a date smaller than a year. Concern for identifiers leads to the obvious step of *deidentifying* the source data, that is, removing identifiers.
2. *Record linkage possible.* Beyond situation 1 of records with univariate identifiers, a data snooper may be able to find attributes in common between the source data and some other database that contains personally identified records. We call such a database an *identity linker file*. Record linkage techniques have become sophisticated and inexpensive and identity linker files are now often easily available to a data snooper. Therefore, confidentiality protection cannot be assured simply through deidentification (Winkler, 1998). To anonymize a record requires more elaborate SDL methods.

There are two ways that a DSO can produce restricted data from source data—the use of disclosure limited masks, introduced above, and the production of synthetic data:

Disclosure-limiting masks. Transform the source data to get a product with low disclosure risk whilst maintaining data utility. Quite general in scope, these masks include perturbations of the original data as well as altering the level of detail or releasing only a sample of the source data (Duncan and Pearson, 1991; Little, 1993; Domingo-Ferrer and Torra, 2001).

Synthetic data. Use the source data to estimate a statistical model (the *synthesizer*) from which simulated data are generated (Rubin, 1993). Only the simulated data—the synthetic data—are released to data users. An alternative term that might better convey its ultimate use is *virtual data*.

Masking and generating synthetic data may be used in combination. For example, in conjunction with overall masking of the source data, suppress-and-reimpute methods suppress certain values and build a synthesizer to simulate replacement values that are *locally* synthetic.

As Fig. 5.4 shows, masking data works directly from the source data, so the masked data have values conditional—deterministically or stochastically—on the source data. Synthetic data, on the other hand, have an intermediate step where the source data are used to build a *synthesizer*, a probability model statistically estimated from the source data. The synthesizer is then used to generate the synthetic data.

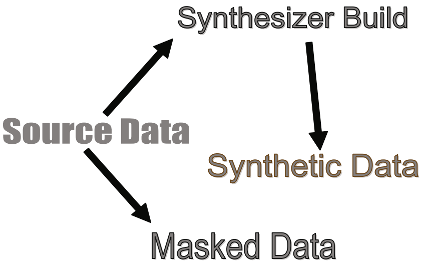


Fig. 5.4 Generating restricted data: masked or synthetic

5.4.1 *In Order to Limit Disclosure, What Shall We Mask?*

Consider Fig. 5.5 in which we partition the attributes in a data record into those considered key variables (or quasi-identifiers) and those considered target variables

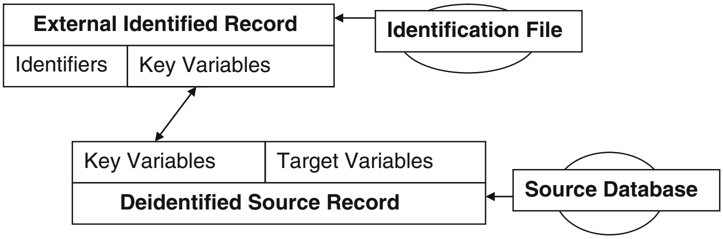


Fig. 5.5 Matching an identity linker file to the source data: linkage through key variables

for a data snooper. The key variables are the entry path to the sensitive variables in that they link to an identified external database. Typically, basic demographic variables, like age, sex, race, occupation, and place of residence are considered to be key variables. As described in Chapter 3, the Data Environment Analysis Service in the United Kingdom is attempting to systematize the identification and classification of key variables, but typically at present "common sense" is the main guide to which variables to consider as key. Target variables are often sensitive and might include income, sexual practices, political beliefs, medical information. Data masking can take either or both of two paths:

1. Mask the key variables making it harder to link.
2. Mask the sensitive variables making them of less use to a data snooper.

Although an overgeneralization, the first path is often more effective in complicating the task of a data snooper—so reducing disclosure risk—yet at the same time leaving most of the interesting research variables unchanged—so maintaining data utility. Standard practice, therefore, is to mask key variables.

5.5 Matrix Masking

In matrix masking we take the source data of n records each with the values of p different attributes to be structured in a rectangular array as a matrix $\mathbf{X}$ with n rows and p columns. To illustrate such source data, consider the Florida Youth Substance Abuse Survey (FYSAS).[15] The survey obtained $n = 7962$ validated records, with each record providing information on a student in grades 6–12. The number of attributes p was large, because it included many substances that can be abused, notably by prevalence, alcohol, cigarettes, and marijuana, but also cocaine, mushrooms, crack, and steroids, among others. Further study attributes included grade level, sex, perceptions of harm, school days skipped, and instances of violent behavior. Note that many of these attributes are sensitive, and so ethical practice requires assurances of confidentiality. Also, without such assurances, many students would not participate or would not provide accurate information if they did.

Like the ways masks at the Carnival of Venice conceal the identity of revelers, disclosure-limiting masks applied to source data can be classified into four categories according to what the masks do:

1. cover up features (suppress data)
2. distort features (perturb data)
3. partially reveal features (sample data)
4. mix into an indistinguishable group (aggregate data).

[15] www.dcf.state.fl.us/mentalhealth/publications/fysas/.

Based on this idea, a classification of disclosure-limiting masks can be laid out according to how they transform data. Further, disclosure-limiting transformations in each of these categories can be represented as matrix masks (Duncan and Pearson, 1991), which have the following structure:

$$\mathbf{Y} = \mathbf{AXB} + \mathbf{C}$$

The source data are the matrix **X** and the masked data are the matrix **Y**. The transformation takes place through

1. *pre-multiplication* of **X** by the matrix **A**, which modifies the rows of **X** (so the data records)
2. *post-multiplication* of **X** by the matrix **B**, which modifies the columns of **X** (so the data attributes)
3. *addition* of the matrix **C**, which directly displaces the values in **X**

In general, the matrices **A**, **B**, and **C** can depend on the values of **X**. Further, any of the three matrices may be stochastic. The DSO specifies the matrices **A**, **B**, and **C**. This specification will impact on both disclosure risk and data utility. We now discuss the four processes for producing restricted data and illustrate each with a matrix mask.

5.6 Masking Through Suppression

A *suppression* is a denial of data instances, so it conceals certain features of the data set (see, for example, Sweeney (2002)). In general, a DSO may choose to suppress data features because of concerns for statistical reliability (say, sample size or survey design questions), data quality (say, worries about respondent exaggeration), or indeed confidentiality. The properties of suppression are, of course, the same regardless of the motivation for using it, but our concern is with suppression conducted in order to reduce disclosure risk. In Chapter 4 we saw how this worked with cell suppression for tables, where for confidentiality reasons the values of particular cells were blocked from user view. With microdata, suppression can take either or both of two forms:

1. *Record suppression.* Deleting data records, for example, dropping all data records for those who have household incomes greater than $200,000 annually.
2. *Attribute suppression.* Deleting attribute values, for example, dropping place of birth because it could be used as a key variable in linking to a publicly available identified record.

Both suppression of a complete record or suppression of a complete attribute can be easily represented as matrix masks:

1. *Record suppression.* Pre-multiply the data matrix **X** by a matrix **A** constructed by dropping from the $n \times n$ identity matrix **I** each row i corresponding to a record i to be suppressed.

2. *Attribute suppression.* Post-multiply the data matrix **X** by a matrix **B** constructed by dropping from the $p \times p$ identity matrix **I** each column j corresponding to an attribute j to be suppressed.

There are three cases where disclosure risk is especially high. The first is where records include longitudinal data, that is, information on an attribute at two or more time points. The second is where records provide geographical specificity, for example, in which block a subject resides. The third is where records have a hierarchical structure, for example, in an educational data setting students within classes, within schools, and in household level data where all members of households are sampled and characteristics of each are recorded.

Longitudinal suppression breaks the links between entity values at different time points. Thus only cross-sectional data would be available at each point in time.

Geographic area suppression typically removes all or some attribute data for geographic areas below a specified population size. For the 2001 Canadian Census, for example, no data were released for areas below a population size of 40. No census data were released for six-character (FSA-LDU) postal codes, geo-coded areas, and custom areas built from the block, block-face, or LDU levels below a population size of 100. In Canada, FSA (Forward Sortation Area) is the first 3 digits of the Postal Code (e.g. M5W) comprising 3000–15,000 Households (average 5000), urban codes A1 to 9A, rural codes A0A; LDU (Local Delivery Unit) is the last 3 digits of the Postal Code (e.g. 1E6, 4V8), used to locate communities within a rural FSA or city block or apartment within an urban Postal Walk. The usual approach to limit the risk of re-identification stemming from geographical information, as discussed by Willenborg and de Waal (2001), is to release according to a broader geographical classification. Such a broader classification usually follows fixed administrative boundaries and therefore does not take account of the particular geographical characteristics of the phenomena under study. An alternative to this was suggested by Franconi and Stander (2002). They employed models to extract information from the original data about the geographical distribution of some of the variables and used this information to inform the broader geographical classification. The same authors have extended in Franconi and Stander (2003) this model-based approach by allowing the model to take account of the spatial configuration underlying the geographical information. In particular, they make use of the neighborhood structure of the geographical areas by employing a conditional autoregressive scheme; see Besag (1974) and Besag et al. (1991). The newer model is flexible and was motivated by the literature on the hierarchical Bayesian modeling of relative risk; see Bernardinelli and Montomoli (1992) and Mollié (1996) for a review. Hierarchical suppression can remove the linkage between respondents, for example in educational data, removing what class and what school a student is in.

5.7 Local Suppression

More complicated suppression patterns are possible that cannot be represented as pre- and post-multiplication. One situation of this sort is *local suppression.* Local suppression suppresses certain values of individual attributes (Willenborg and de Waal, 2001). The aim is to reduce the set of records in which there are only a few combinations of particular values. When local suppression is applied, one or more values in a risky combination are suppressed and replaced by a "missing" value. For instance, the combination, "Place of residence = Los Angeles", "Sex = Female", and "Occupation = Civil Engineer" might have high disclosure risk, which could be reduced by suppressing the value of "Occupation", since the number of females in Los Angeles is high. Of course the resulting combination "Place of residence = Los Angeles", "Sex = Female", and "Occupation = missing". may not be of much value to many data users, certainly say, to a labor economist. So instead of suppressing the value of "Occupation", we could instead suppress the value of one of the other attributes, say "Place of residence" instead of the value of "Occupation". You can certainly imagine other statistical disclosure-limiting schemes that might retain more data utility, perhaps recoding to aggregate all engineers. We will explore a number of such SDL schemes later, but first more about suppression.

Suppression is local if it is applied only to particular values of an attribute. Hence if the value of "Occupation" is suppressed in a particular record, then it is not necessarily the case that the value of "Occupation" will be suppressed in another record. This flexibility in selecting the values that are to be suppressed, and so not having to suppress all of the values of an attribute, can lower the number of suppressions needed to adequately lower disclosure risk.

An alternative to suppression is *perturbation*, in which the DSO distorts attribute values in data records. Because it is generally accepted that probabilistic methods of distortion have lower disclosure risk than deterministic ones, perturbation methods typically involve stochastic changes to attribute values. We discuss the two most important perturbation methods, adding noise (Section 5.8) and data swapping (Section 5.9).

5.8 Noise Addition

Perturbation of an attribute value x by noise addition adds a random variable ε, so the masked value y has the form $y = x + \varepsilon$. To avoid bias, the mean of ε is taken to be zero. Generally this method is used when x takes on values over some continuum, but it could be used when x is discrete as well by imposing restrictions on the values of ε (see Section 4.3.2).

The extent of perturbation can be raised by increasing the variance of ε. Referring to Fig. 5.5, such noise is generally added to key variables. This perturbation is intended to stymie any attempt by a data snooper to link to an identity linker database. Adding noise to a sensitive variable is generally less successful and so less used. The reason for this is that the level of noise required to adequately fuzz the values must have such a high variance that the values lose their utility for data analysis.

The practice of adding noise has been examined as a disclosure limitation device by several authors. These include, in the case of a univariate x (so it encompasses just a single attribute), Brand (2002a, b), Domingo-Ferrer et al. (2004), Spruill (1983), Paass (1988), and Duncan and Mukherjee (1991, 2000), and in the case of a multivariate $\mathbf{x}$ (so it encompasses two or more attributes), Sullivan and Fuller (1989) and Fuller (1993). Also see Kim and Winkler (1995, 2001), Muralidhar et al. (1999), Mera (1997), Tendick (1991), and Tendick and Matloff (1994) for further investigation of this method.

To further develop the concepts of this section in a mathematical way that gives appropriate forms for implementation, we need to have some notation and some assumptions. The material that follows requires additional mathematical and statistical background and may be skipped by a reader who is primarily interested in the concepts. The key idea is that although appropriate forms of noise addition can keep the means and covariances of the masked data unchanged from that of the source data, variances and correlation coefficients will be changed. In particular, correlation coefficients in the masked data will appear closer to zero than they are in the source data.

Take the source data to be represented as the $n \times p$ matrix $\underset{n\times p}{\mathbf{X}} = \begin{bmatrix} \underset{1\times p}{\mathbf{X'}_1} \\ \vdots \\ \underset{1\times p}{\mathbf{X'}_n} \end{bmatrix}$, where each of the n rows correspond to a particular record. These records are taken to be independently and identically distributed so the p attributes each have mean vector $\boldsymbol{\mu}$ and cariance - convariance matrix $\boldsymbol{\Sigma}$. We write $\mathbf{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In matrix notation take the masked microdata as $\underset{n\times p}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}} + \underset{n\times p}{\boldsymbol{\varepsilon}}$, where the rows of the noise $\underset{p\times 1}{\boldsymbol{\varepsilon}} \sim \left(\underset{p\times 1}{\mathbf{0}}, \lambda^2 \underset{p\times p}{\boldsymbol{\Sigma}}\right), i = 1, \ldots, n$, are uncorrelated.

A simplifying fact for data analysis using the masked data $\mathbf{Y}$ is that this additive perturbation preserves both means and covariances, that is,

$$E(\underset{n\times p}{\mathbf{Y}}) = E(\underset{n\times p}{\mathbf{X}}) + E(\underset{n\times p}{\varepsilon}) = E(\underset{n\times p}{\mathbf{X}}) = \underset{n\times p}{\mathbf{M}}$$

and for each pair of column vectors of $\underset{n\times p}{\mathbf{Y}}$,

$$\mathrm{Cov}(\underset{n\times 1}{\mathbf{Y}_i}, \underset{n\times 1}{\mathbf{Y}_j}) = \mathrm{Cov}(\underset{n\times 1}{\mathbf{X}_i}, \underset{n\times 1}{\mathbf{X}_j}).$$

Unfortunately for data analysis, neither variances nor correlation coefficients are preserved:

$$\mathrm{Var}(\underset{p\times 1}{\mathbf{Y}_j}) = \mathrm{Var}(\underset{p\times 1}{\mathbf{X}_j}) + \mathrm{Var}(\underset{p\times 1}{\boldsymbol{\varepsilon}_j}) = (1+\lambda^2)\Sigma, j = 1, \ldots, n;$$

$$\text{so, } \rho(Y_r, Y_s) = \left(\frac{1}{1+\lambda^2}\right) \rho(X_r, X_s).$$

Note in particular that the correlation coefficients are attenuated and thus nonzero correlation coefficients in the source data with the X's will appear closer to zero in the masked data, the Y's.

5.9 Data Swapping

Data swapping switches certain fields in a record with corresponding fields in another record. In the statistics literature, swapping observations "at random" was proposed by Dalenius and Reiss (1978) and developed by Dalenius and Reiss (1982), both for the case of categorical data. Reiss (1984) and Dalenius (1988) describe an extension to continuous (numerical) data. Also see Moore (1996), Reiss et al. (1982), and Takemura (2002).

Earlier, in the computer science literature, Conway and Strip (1976) suggested a method they called *value disassociation*—a database query for the value of a record field yields the value of the same field in some other record. They propose that if a field for record i is included in a database query that it be replaced by the same field of another record j. The record j would be selected at random and with replacement. Data swapping is also referred to as *multidimensional transformation* (Schlörer, 1981) and *data switching* (Navarro et al., 1988). A review of data swapping is Fienberg and McIntyre (2004).

Sanil et al. (2003) describe an implementation, called NISS (National Institute of Statistical Sciences) WebSwap, of a web service for data swapping. As an online service, NISS WebSwap swaps according to user specifications. The software is available from the NISS web-page, www.niss.org. Using the R-U confidentiality map framework first proposed by Duncan and Fienberg (1999) and developed by Duncan et al. (2001) and consistent with the viewpoint of Zaslavsky and Horton (1998), Gomatam et al. (2005) discuss applications to categorical data, with an assessment of risk through the proportion of unswapped records and an assessment of date utility through a Hellinger distance distortion measure; Gomatam et al. (2005) examine an instantiation incorporated into NISS WebSwap.

Data swapping transforms the source data matrix $\mathbf{X}$ of n records (rows), each with p attributes (columns), into another $n \times p$ matrix $\mathbf{M}$. The data product to be released is $\mathbf{M}$ or some statistics based on $\mathbf{M}$. Data swapping is intended to lower disclosure risk whilst explicitly maintaining, or approximately maintaining, certain statistics calculated from $\mathbf{M}$. In this sense, $\mathbf{M}$ is equivalent or near equivalent to $\mathbf{X}$, and for some applications would have high data utility.

Implementation costs of data swapping have been regarded as high (Eurostat, 1996, p. 63). A computational disadvantage of data swapping is that it is a global method that requires knowledge of values in records throughout the database (Agrawal and Srikant, 2000), whereas other methods, say the usual incarnation of noise addition (e.g., Duncan and Mukherjee, 1991), are local methods that act on and depend on a particular record value.

5.9.1 Implementations of Data Swapping

The US Census Bureau used data swapping as part of the *Confidentiality Edit* in preparing public-use files from the 1990 Decennial Census. Staff exchanged an unspecified fraction of household records in different census blocks whilst holding constant a certain subset of attribute values. Preliminary to this implementation, Navarro, Flores-Baez and Thompson (1988) and Griffin et al. (1989) described an empirical study using 1980 census data for the state of New Jersey (also see Subcommittee on Disclosure Avoidance Techniques, 1994). Geography, specifically census blocks, provided the *swapping attribute*, so records were swapped from one census block to another. Matching of records was done on household size, race, Hispanic origin, and number of persons aged 18 and over. These attributes were the *swapping keys* (see Domingo-Ferrer and Torra (2003)).

Data swapping introduces uncertainty in the mind of the data snooper about whether the linkage is correct. This uncertainty deters the data snooper from making an identification and hence lowers disclosure risk. The matching on swapping key attributes by the DSO guarantees that statistics based on the swapping keys are invariant under data swapping. Thus, in this respect, data utility is not reduced. No such guarantees apply to statistics based on location (the swapping attribute) or the other variables (the protected variables). Willenborg and de Waal (1996, p. 16) and Fienberg et al. (1996) provide further discussion. The US Census Bureau applied data swapping for public releases from the 2000 Census (Zayatz et al., 1999), including web-based releases through the American FactFinder system (Zayatz and Rowland, 1999).

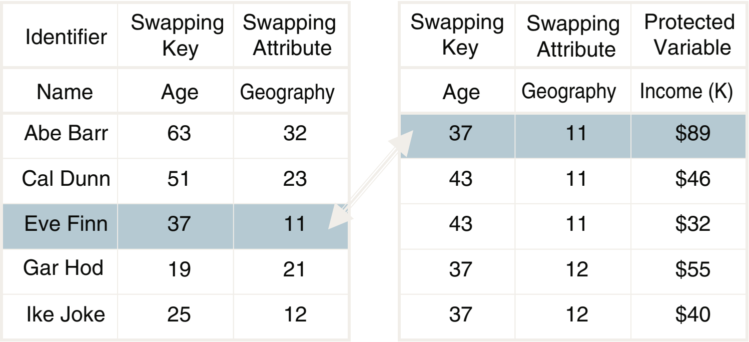
Data swapping is also sometimes coupled with other disclosure limitation methods, for example, top coding. Kim and Winkler (1995) provide a disclosure limitation methodology (the Kim–Winkler approach) that couples an additive noise approach with a secondary use of data swapping. The methodology is applied to the task of producing a public-use database that contains Current Population Survey (CPS) data and income data from the Internal Revenue Service (IRS) 1040 Form.

5.9.1.1 An Example

Consider a simple illustration, inspired by a survey of US doctorates (Duncan et al., 1989). The data on physicists employed by academic institutions comprise age, university, and salary.

In all applications, domain knowledge is needed to specify which variables are to be considered key. With the doctorate survey it is reasonable that a data snooper has access to the name and age of some physicist. In swapping records, matching would be done on age, the swapping key. Records would be swapped from one university to another so the swapping attribute is university. In swapping from one university, say University 11, another university, say University 12, would be chosen to have similar characteristics with respect to the protected variables—in this case salary. The labeling of the universities is deliberate: universities with the same first index have approximately the same distribution of salaries. There is flexibility in defining measures of closeness, separately for the swapping key and the swapping attribute. Here age is measured to the nearest year and an exact match on this variable is possible. The closeness of two universities might well be based on published information about their salary distribution—say through the American Association of University Professors (AAUP) Annual Faculty Compensation Survey.[16]

In this survey of doctorates example, a data snooper might have knowledge of the names and ages of physicists at particular universities. For example, the data snooper may have access to the database on the left which could be linked to the subject database on the right. In this application, "geography" corresponds to "university."



Identifier	Swapping Key	Swapping Attribute
Name	Age	Geography
Abe Barr	63	32
Cal Dunn	51	23
Eve Finn	37	11
Gar Hod	19	21
Ike Joke	25	12

Swapping Key	Swapping Attribute	Protected Variable
Age	Geography	Income (K)
37	11	$89
43	11	$46
43	11	$32
37	12	$55
37	12	$40

Because the 37-year-old in University 11 is unique on the swapping key, the data snooper could identify this record and learn that this person's salary was $89,000 per year. Under data swapping, the protected variables for this individual might be swapped with those of the 37-year-old in University 12. The publicly known possibility that this swapping has been done is intended to (1) dissuade the data snooper from attempting to make such a linkage, and (2) lower the public credibility of any claim by the data snooper of a valid identification.

5.9.2 A Protocol for Data Swapping

The previous illustration and the logic of practice suggest the following definitions:

Swapping candidate. A swapping candidate is a record having characteristics that suggest it for swapping. In some data files, all records may be swapping candidates. More often—particularly with large data files—only certain records are swapping candidates. These records may be suggested by the sensitive nature of the attribute values (say, certain psychiatric diagnoses in a medical record) or to the ease with which they may be identified. Identification may be easy if the data snooper has access to an identified

[16] (chronicle.com/stats/aaup/)

database containing attributes that overlap with that of the subject database and some records are outlying because of unusual combinations of attributes (the 80-year-old marathon runner). With multiple attributes—such as age, sex for all members of a household—available to the data snooper, a high proportion of records are likely unique in even quite large populations (see Bethlehem et al. (1990) for an illustration of the Netherlands).

Swapping key. The attributes (age in our example) which are used to match records for swapping are called swapping keys.

Swapping attribute. An attribute over which swapping is to occur (university in our example, but often geographical units, such as census tract, or political units, like counties).

Swapping partner. A record in a counterpart equivalent, in some sense, to a swapping candidate.

Data swap. The exchange of record values between a swapping candidate and its swapping partner. Swapping occurs if a suitable swapping partner can be found and a probabilistic mechanism marks it to be swapped.

Under data swapping, both disclosure risk and data utility are affected by several factors:

1. Criteria for selecting swapping candidates
2. Specification of a swapping key
3. Similarity among swapping attributes
4. Similarity of entities within swapping attributes
5. Swapping rate
6. Knowledge, by the data snooper and the data user, of the swapping protocol, including the swapping rate used

To provide a basis for discussion of these factors, consider the following protocol for implementing data swapping that we will call the *Basic Protocol*:

1. Any attribute (age, in our example) is in the swapping key when the DSO believes that a data snooper may well have knowledge of the attribute values.
2. A record is a *potential swapping partner* if it has similar values on the key attributes and has not previously been swapped. If there is more than one potential swapping partner, select one at random. Choosing from the set of potential swapping partners at random is equivalent to initially randomizing the order of records and selecting the first potential swapping partner.
3. Protected attribute values are stochastically swapped between those of a swapping candidate and those of its chosen potential swapping partner, independently with probability π. The value of π would be determined by the DSO. Intuitively, the higher the probability of a swap, the lower the disclosure risk, but also the lower the data utility.
4. The swapping rate π is not revealed to either the data snooper or data user.

Although this Basic Protocol for implementing data swapping appears reasonable, there are many other possible ways a swapping protocol could be set up. For example, it may be appropriate to give swapping candidates different priorities for confidentiality protection. In a medical context, for example, one diagnosis (AIDS) may require that the record be given more protection than another diagnosis (rotator cuff tendonitis). Greater sensitivities of records suggest higher swapping probabilities. Also, the swapping rate π might be revealed. This would increase both disclosure risk and data utility.

Consider now data utility under the Basic Protocol. Quite generally, we can address the question of when there is no inferential loss under data swapping: as noted before, for an inference population that is conceptually infinite there is no inferential loss when the conditional distribution of the protected variables given the key variables is the same for each swapping attribute value. Thus, if inference is to be done conditionally on the key variables, the inferences are not distorted if observations are exchangeable. So there is no loss in data utility when the values to be swapped are exchangeable. Motivated by the fact that in many real instances the values could not be considered exchangeable, we address in Chapter 6 a quantification of the loss in data utility under data swapping.

5.10 Masking Through Sampling

In *masking through sampling* the masked data are a sample of the source data. If the source data are themselves a sample, the data may be considered self-masked, but just the fact that the data are a sample may not result in disclosure risk that is sufficiently low. In that case, subsampling and/or the use of other disclosure limitation methods may also be required.

Sampling is often used in preparing public-use microdata. Statistics Canada, for example, published the 2001 Public Use Microdata File (PUMF) with approximately 3% of the Canadian population. In addition, data for small geographic areas are not available in these files. Information is provided only for selected census metropolitan areas, selected census sub-divisions, the provinces, and the territories. The detail of some sensitive variables was reduced for the Atlantic region. Some of the values of sensitive variables were suppressed because their combination could have been used to identify a person, a family or a household. Finally, income variables were top and bottom coded. For households in the Atlantic Region, for example, income was bottom coded below a negative $30,000 and top coded above a positive $120,000.

How much is disclosure risk lowered through sampling? Firstly, and perhaps obviously, if a population unit is not in the sample then no information about its identity can be disclosed about it. Secondly, the intruder cannot be certain whether any linkage between a population unit and a data unit is correct. It might be the population unit's statistical twin that has been sampled, and linked to, but not the population unit itself. As discussed in Chapter 3, how much uncertainty is created depends on how likely a unit linked to an identifier in the sample is unique in the population (see Samuels (1998), Elliot et al. (1998) and Skinner and Elliot (2002) for further discussion of this).

5.11 Masking Through Aggregation

Aggregation as a masking method involves statistically combining attribute values for some data records or statistically combining values of different attributes (or even both). Essentially in most implementations it is a coarsening method. Aggregation methods include *global recoding* and *topcoding*.

5.11.1 Global Recoding

Global recoding combines several categories of an attribute into less-specific categories. Literature on global recoding includes Bayardo and Agrawal (2005), DeWaal and Willenborg (1995, 1999). In common applications of global *recoding*, several categories of an attribute are collapsed into a single one. For example, the US Bureau of Labor Statistics conducted the National Compensation Survey where the source data had an attribute, "Occupation" that included both chemical engineers and civil engineers. For some purposes they, along with petroleum engineers and mechanical engineers and some other categories, are recoded as, simply, engineers. This lessens the chances of an exact match on key variables to an identified external database. Another attribute that might be subject to global recoding would be "Place of Residence," where both Albuquerque, New Mexico and Denver, Colorado might be placed in the Mountain region. The recoding in these examples is global in that *every* record is placed in some broader category. Global recoding is applied to the whole file, not only to records that have especially high disclosure risk. This eases statistical analysis because it gives uniform categories for each attribute. Suppose, for instance, that we recode the "Occupation" attribute so all engineers are lumped together. Though the combination "Occupation = civil engineer," "Place of Residence = Los Angeles" and "Occupation = chemical engineer," "Place of Residence = Los Angeles" might not have high disclosure risk, the only published cases would involve, "Occupation = engineer," "Place of Residence = Los Angeles". Sometimes global recoding decisions are made on the basis of the dependency structure across several key variables. For example, for the general release microdata, in the 2001 UK census, ONS used much coarser codes for the ages of people of working age than for children and the retired. This was because the combination of the detailed occupation variables, that researchers requested on this data set, and age increased the disclosure risk beyond what ONS considered acceptable. This problem did not apply to children and those of retirement age and so more detail was possible for the age variable for those data units.

5.11.2 Topcoding

Topcoding is a specific example of global recoding. It typically involves replacing all attributes values above a certain threshold with either just a notation to that effect or some summary statistic such as the conditional (on being above the threshold) mean or median. Similarly, bottom coding replaces values below a certain threshold. Age is a variable which is routinely subject to top coding (as the number of population units at each age decreases as age increases thus increasing the disclosure risk).

5.12 Microaggregation

The *microaggregation* process is thus: Identify any attributes, such as income levels, that pose disclosure risk. Partition records into clusters according to these attributes. An aggregate value, such as an average, mode, or median, is assigned to each of the values of the attributes identified as posing disclosure risk. The rationale is to ensure each record value corresponds to at least some minimal number of original individuals where no one or two individuals dominate.

Relevant literature on microaggregation includes Dandekar et al. (2002), Defays and Nanopoulos (1993), Defays and Anwar (1998), Domingo-Ferrer (2001), Domingo-Ferrer and Mateo-Sanz (2001), Domingo-Ferrer et al. (2002), Mateo-Sanz and Domingo-Ferrer (1999), Oganian and Domingo-Ferrer (2001), Sande (2002), Torra (2004) and Valls et al. (2002).

In early work, Defays and Nanopoulos (1993) suggested forming clusters of the same size, but more recent research, specifically Domingo-Ferrer and Mateo-Sanz (2002), Hansen and Mukherjee (2003), Domingo-Ferrer and Torra (2005), and Laszlo and Mukherjee (2005) suggests advantages to clusters of variable sizes.

5.13 Synthetic Microdata

The methods described so far have masked the source data. These are called *data-conditioned* methods by Duncan and Fienberg (1999). Another approach is conceptually familiar to statisticians: consider the original data to be a realization according to some statistical model; replace the original data with samples (the synthetic data) generated according to the model. Synthetic data consist of records of individual synthetic units rather than records the data stewardship organization holds for actual units.

In initially proposing the use of synthetic microdata, Rubin (1993) asserts that the possibility of identity disclosure can be eliminated through the dissemination of synthetic data. This is because the synthetic data carry no direct functional link between the original data and the disseminated data. So whilst there can be substantial identity disclosure risk with (inadequately) masked data, identity disclosure is, in most cases, not possible with the release of synthetic data. However, the release of synthetic data may still involve risk of attribute disclosure (Fienberg et al., 1998). Domingo-Ferrer and Torra (2005) argue that useful synthetic data may for certain records be sufficiently similar to the source data as to be disclosive. However, this is a debatable position; as any inferences that can be made about a population unit from synthetic data are effectively inferences that can be made from the statistical model that generated that data. If we are to adopt the position that such inferences can be considered disclosive, then we must, in principle at least, extend the argument to any statistical model generated from any data set which would seem counterintuitive and certainly a position that most data analysts would want to avoid! We would argue that synthetic data create a situation where far from being *uncertain* of any correspondence between a given data unit and a given population unit the intruder will be *certain that correspondence is not real*. This is because the data unit has not been drawn from the source microdata. Therefore, the intruder who uses synthetic data to generate plausible inferences about particular attribute values for particular data units is not engaged in statistical disclosure at all, but in what marketing people call profiling a quite different and so far legitimate activity.

Rubin (1993, 1996) cogently argues that the release of synthetic data has advantages over other data dissemination strategies, because:

1. Masked data can require special software for its proper analysis for each combination of analysis × masking method × database type (Fuller, 1993; Little, 1993).
2. Release of aggregates, for example, summary statistics or tables, is inadequate due to the difficulty in contemplating at the data-release stage what analysts might like to do with the data.
3. Mechanisms for the release of microdata under restricted access conditions, for example, user-specific administrative controls, can never fully satisfy researcher demands.

An early application of the synthetic data approach is by Kennickell (1999). More recently, as Wu and Abowd (2008) notes, “The Longitudinal Employer-Household Dynamics Program at the U.S. Census Bureau has developed several synthetic data products including OnTheMap and a partially synthetic version of the Survey of Income and Program Participation linked to Social Security Administration and Internal Revenue Service data.”

The methodology for the release of synthetic data is simple in concept, but complex in implementation. Conceptually, the agency would use the original data X to estimate the underlying cumulative distribution F_X and generate and synthetic samples from $\hat{F}_X$. As suggested by Rubin (1993), in order to obtain proper variance estimates of various parameters, multiple synthetic samples can be released. The complexity in these techniques arises in determining how to estimate F_X, the size sample to generate from $\hat{F}_X$, and how many samples to generate.

Further, the purpose of this model is not the usual prediction, control, or scientific understanding that argues for parsimony through Occam’s Razor. Instead, its purpose is to generate synthetic data useful to a wide range of users. The agency must recognize uncertainty in both model form and the values of model parameters. This argues for the relevance of hierarchical and mixture models to generate the synthetic data.

Synthetic data is a vigorously pursued research area. Little (1993) proposed replacing only the "sensitive values" with synthesized data, yielding "partially synthetic" microdata. Further contributions to research on synthetic data include Abowd and Woodcock (2002, 2004), Raghunathan (2003), Raghunathan et al. (2003), Raghunathan and Rubin (2000), Reiter (2002, 2003, 2005a, c, d), and Thibaudeau and Winkler (2002). Machanavajjhala et al. (2008) generated synthetic data in the context of geospatial data (discussed further in Section 8.5.1) and concluded that future research is needed on methods for incorporating outlier identification, suppression, and modeling for this mapping application.

5.14 Concluding Thoughts

Increasingly, analysts want microdata. For example, in the United States a BlueCross/BlueShield Association might seek access to certain personal healthcare utilization data, not for the obvious administrative reasons, but for research purposes that might lead to better practices. Quite possibly, these utilization data would be microdata on subscribers to particular health management organizations (HMOs). In addition, Blue Cross/Blue Shield would presumably want three other personal characteristics:

1. subscriber demographic attributes
2. HMO corporate structure and co-payment requirements
3. subscriber's employer through its nature of business, size, and health insurance alternatives made available to employees.

In Section 5.1 we outlined why users need such data. But as we argued in Section 5.2, a DSO has ethical, pragmatic, and legal motivations to provide confidentiality protections. Also, we argued in Section 5.3 that microdata are especially vulnerable to confidentiality attack. Providing microdata runs into confidentiality constraints that are not easy to satisfy. Because of the obvious richness of many records, deidentification—stripping apparent identifiers such as name, SSN, email address—does not prevent reidentification through linkage with external databases that included identifiers (notably marketing databases but increasingly data collected for a wide range of governmental, social, and commercial purposes). With such reidentification, the data snooper would have successfully attacked the database and there would have been an inferential disclosure that compromised confidentiality. Examples like these indicate the need for SDL methods for microdata to be used. Sections 5.4–5.12 explored various masking methods, including matrix masking, suppression, noise addition, data swapping, masking through sampling, global recoding, topcoding, and microaggregation. Finally, Section 5.13 introduced synthetic data as data stochastically generated from a model inferred from the source data. With the SDL tools discussed in this chapter we have the means to both provide and protect microdata.

Chapter 6
Disclosure Risk and Data Utility

As we have repeatedly argued, DSOs fulfill their stewardship responsibilities by resolving the tension between ensuring confidentiality and providing access (Duncan et al., 1993; Kooiman et al., 1999; Marsh et al., 1991). Data stewardship, therefore, requires disseminating data products that both (1) protect confidentiality—so get disclosure risk R low by providing *safe* data and (2) keep data utility U high by providing data products that are *analytically valid*. In other words, the problem of protecting data is bi-criteria. This opens the question of how to balance the two criteria. Answering this requires that we know how R and U affect each other.

We begin this chapter by showing several ways disclosure risk R and data utility U can be assessed. With the pairs (R, U) calculated for different levels of disclosure limitation (as would be the case, for example, with different levels of the variance in noise addition), we can examine how R and U are traded off by a given SDL procedure. This gives us an *R-U confidentiality map* (Duncan et al., 2001). Also, comparing such maps for different SDL techniques can provide a basis for choosing among them (for example, comparing the R-U confidentiality map for noise addition with that for data swapping).

6.1 Basics of Disclosure Risk and Data Utility

As discussed in Chapters 1 and 2, when a data snooper compromises confidentiality a disclosure takes place (Elliot and Dale, 1999). The disclosure could involve the data snooper learning who the data subject is—*identity disclosure*. Or it could involve the data snooper learning more than the snooper should about a data subject—*attribute disclosure* (Duncan and Lambert, 1989). To prepare a data product that is safe from attack, a DSO must restrict the data by employing SDL methods such as those described in Chapters 4 and 5. These methods can be effective in lowering the disclosure risk R (the measurement of which was described in Chapter 3).

But after lowering R what happens to the data utility U? To see the impact of transforming source data into masked data, consider one easily interpreted and

G.T. Duncan et al., *Statistical Confidentiality*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-7802-8_6,

implemented SDL method: coarsen the data by creating bins according to ranges of attribute values and counting the number of occurrences in each. For example, recode income in increments of $5000 and release a table giving, say, how many earned between $60,000 and $65,000. The coarsening of a key variable lowers disclosure risk by making reidentification less likely. Coarsening of a target variable lowers the payoff to the data snooper of an attribute disclosure. For example, in the above example only income group is disclosed. This can only lower the incentive for a data snooper.

Thus, coarsening lowers the disclosure risk R, which is fine for data protection. But from the perspective of the data user coarsening makes things worse. In our example, regression of income on education becomes more difficult for the user and less informative. Or for a more dramatic example, coarsen a male/female dichotomy, and you have lost the sex attribute entirely, a sad situation indeed. Many data users, especially those who can make the most important research and policy contributions and who command the latest computer technology, require and can use data of high resolution. Often they will not be satisfied with coarsened data because the data utility U would be too low for their purposes.

As we discussed in Chapters 4 and 5, SDL techniques provide classes of data transformations that lower disclosure risk. There are a cornucopia of available SDL methods, each with different impacts on data utility and disclosure risk. This complicates the DSO's task. Which SDL method should the DSO use? Are there criteria for choosing?

To begin answering such questions, we show some ways of assessing the simultaneous impact on disclosure risk R and data utility U of implementing an SDL technique. Recall that as a measure of statistical disclosure risk, R is a numerical assessment of the risk of disclosure following dissemination of the data. As a measure of data utility, U is a numerical assessment of the usefulness of the released data for legitimate purposes.

6.1.1 Choosing the Parameter Values of an SDL Method

Consider a single SDL method, such as data swapping. The disclosure risk R and the data utility U depend on the parameters set for the method, in this case the swapping rate. In general, there are two widely used approaches to selecting the "best" parameter values for the SDL method:

1. Maximize the utility among those that have disclosure risk below a fixed threshold ρ; that is, choose parameters to obtain $\max\{\mathrm{U} : \mathrm{R} \leq \rho\}$.
2. Maximize a score that is a weighted average of the R and U measures, that is, choose parameters to obtain $\max\{\lambda\,\mathrm{R} + (1 - \lambda)\mathrm{U}\}$ for some λ.

Approach 1 of maximizing U subject to a constraint on R is the more commonly used. Also it is familiar to all trained in statistics because it is in the same spirit as choosing a hypothesis test by maximizing power subject to the probability of a Type I error being below some fixed level. Approach 1 avoids the necessity with Approach 2 of having to choose and trade off λ and coping with the attendant problem of different scales for R and U. Approach 1 puts primary emphasis on confidentiality, as masked data are only released when the risk of disclosure is under a threshold.

Approach 1 does have complications, however. The threshold, here represented by ρ, may actually be multi-attributed. For instance, Chapter 4 has illustrated different SDL methods to publish tables and ρ can be taken as the collection of all the disclosure limitation levels required by the DSO, for each sensitive cell and each data snooper. Further, in spite of its intuitive appeal and common usage, Approach 1 has been criticized, especially as it may lead to "extreme criteria." Suppose, for example, that both R and U take values in the range (0,100) and that the threshold for safety is $\rho = 9$. Also, you have two alternative data releases, D_1 that produces the pair ($R = 8.999$, $U = 10$), and D_2 that produces the pair ($R = 9.001$, $U = 90$). Based on Approach 1, D_1 should be selected. In this case, however, a DSO might well choose D_2. In fact it is hard to believe that an arbitrary large increase in validity is not worth an infinitesimal decrease in safety. Method 2 could therefore be more appropriate. Nevertheless, the choice of the weight λ should be based on solid theoretical ground (see, for example, Domingo-Ferrer and Torra (2001)). This criticism is applicable to any technique that tries to reduce a multi-objective decision problem into a single-objective decision problem, and indeed the literature on decision making has extensive discussions of this criticism that also apply to SDL. See, for example, Kirkwood (1996), Shlomo (2003), and Trottini (2003).

In order to be able to quantify data utility, we now introduce a variety of metrics.

6.2 Data Utility Metrics

Data utility is familiar to statisticians and researchers because it relates to the value of data in statistical inference. Data utility metrics appear naturally as power of hypothesis tests, mean squared error of estimators, and width of confidence intervals. Agrawal and Srikant (2000) develop a metric for data utility based on the width of the 95% prediction interval. Generally, applying SDL to transform a source data set to a masked data set lowers data utility. This loss may not be evident to all data users, especially those who are not sophisticated researchers, because the quantity of data available may not seem to be diminished. Kamlet et al. (1985) give examples that show how SDL techniques can seriously alter observed relationships among data attributes, and hence hamper data utility. The role of data utility in SDL has been further recognized by Kennickell and Lane (2006), Kim (1986), Marsh et al. (1991), Skinner (1990), and Trottini (2003).

A DSO should make the fact of loss of data utility from SDL evident to the data user and also provide measures of the extent of this loss. Basically this can be measured by an assessment of how accurate would be inferences from the masked data relative to the accuracy of inferences from the source data. But a DSO may find it difficult to pre-specify the uses of its data product. With the particular tasks unclear, the DSO may instead release a data product that does not differ much in its general informational structure from the source data. This implies measuring data utility to be based on *information loss metrics*. The hope is that by minimizing information loss the released data will preserve enough of the information in the source data for it to remain analytically valid and interesting.

There are two complementary methods to assess information loss:

1. Compare the data in the source and the masked data.
2. Compare certain statistics computed on the source and the masked data.

Following Gomatam and Karr (2003), with the first method, based on the discrepancy between the masked data and the original data, we call the information loss metric a *distortion measure*. If we choose the second method, based on the discrepancy between statistics from the source data and the masked data, we call it a *proxy measure.*

To appreciate how information loss metrics can work, consider an example in Shlomo (2003) of a study of the determinants of income. The key variables that a data snooper might typically use for record linkage are categorical demographic and geographic identifiers. Mostly, they are used by researchers and analysts as explanatory variables in regression models. For SDL based on coarsening of the categories there are two evident information loss metrics:

1. The loss of the predictive power of a regression model, as expressed by R^2, where the dependent variable is income and the independent variables are the demographic and geographic variables that are collapsed.
2. The loss in information as the keys get coarser and the categorical variables are collapsed as measured by entropy.

6.3 Direct Measurement of Utility

Although the metric approaches discussed in Sections 6.1 and 6.2 have important advantages in providing general measures that relate to data utility, they all suffer from the same flaw: they do not measure utility directly. That is they do not relate to specific analyses that users intend to or will conduct. Instead, they are an attempt to capture utility through abstraction. Purdam and Elliot (2007) stress that such measures are proxies for the impact of SDL on *analytical power,* which they define as "the ability of the user to draw correct conclusions from a given set of data."

Analytical power is related to, but distinct from, statistical power and consists of two components *analytical completeness* and *analytical validity*. Completeness is a conceptual measure of the ability of the data user to run the analyses that they wish to run. Validity here is the ability of the user to derive the same conclusions from running a given analysis that they would have if they had perfect data. Both of these are affected by other things besides the SDL process. For example, completeness is affected by what data have been collected in the first place and validity by all forms of data divergence (including measurement and data processing error).

However, as Purdam and Elliot (2007) emphasise, it is also clear that analytical completeness and validity are affected by SDL procedures. With the collapsing of categories, analyses that might have been conducted with unrecoded data cannot be (i.e., completeness is reduced). For example, the use of geographical thresholds in microdata sets leads to smaller administration units being grouped together, thereby preventing researchers from effectively using the data set for inferences about those smaller units. On the other hand, any perturbative SDL technique can change a data set to the point where a user reaches a different conclusion than would be derived from the same analysis conducted on the unperturbed data.

In a case study using 2001 UK census microdata, Purdam and Elliot carried out two direct studies of utility. The first study was a survey of researchers with published analyses of the data to establish the effect of 28 possible recodes of the released data on the ability of users to conduct the published analyses (i.e., the impact on completeness). The second study applied the SDL system μ-ARGUS to perturb the data set and then reran some reported analyses to assess whether the same conclusions could be reached after μ-ARGUS had perturbed the data. The results of both studies were of some concern. The recoding led to a reduction in the analyses that could be run. The perturbation led to impact on interpretation in the majority of cases.

With the understanding we have developed about disclosure risk and data utility, we will now examine a systematic approach to examining the tradeoff between R and U.

6.4 The R-U Confidentiality Map

Present practice by DSOs in assessing tradeoffs between disclosure risk and data utility is primarily heuristic. Recommendation 6.2 of the National Academy of Sciences Panel on Confidentiality and Data Access (Duncan et al., 1993, p. 196) essentially advises the development of foundations for such determinations. Approaches to disclosure risk based on a decision-theoretic characterization of the data snooper are developed by Duncan and Lambert (1986, 1989), Lambert (1993), Little (1993), Mokken et al. (1992), and, more comprehensively, for both the data snooper and the DSO by Trottini (2001). The idea is to view the actors as decision makers, who take actions in light of their perceptions of probabilities and consequences:

1. The data snooper can choose (or not) to make identifications and draw inferences on the basis of the released data product.
2. The DSO chooses an SDL method to deter the data snooper.

From this perspective, disclosure risk depends on the decision structure—probabilities and utilities of consequences—of the data snooper and the DSO. There are, however, important complications, (a) and (b), from the usual decision model:

(a) The DSO has only its own perceptions of the decision structure of the data snooper.
(b) The DSO must cope with multiple data snoopers.

The solution to (b) is general—protect against the worst. Although at first thought protecting against the worst seems overly conservative, it is actually reasonable given that a DSO is concerned about a disclosure to *any* data snooper. The solution to (a) is also general—put yourself in the shoes of the data snooper. Implementing both of these solutions reduces the problem back to the decision structure of a single individual.

In this section we demonstrate how the R-U confidentiality map provides a useful analytical framework for DSOs. It allows them to assess tradeoffs between the benefits of providing data products and the risks involved in doing so. With the R-U confidentiality map, the DSO has an analytic tool for systematically examining tradeoffs between value to a client, *data utility*, and vulnerability to a data snooper, *disclosure risk*. A beginning on this task in the form of a basic R-U confidentiality map is presented in Duncan and Fienberg (1999) in the context of tabular data. A further development, again in the context of tabular data, is given in Duncan et al. (2001). In this section we develop R-U confidentiality maps in the context of continuous microdata, along the lines laid out by Duncan et al. (2001). See also Shlomo (2003) for extensions to global recoding, as well as comparisons of disclosure limitation via additive noise and global recoding.

An R-U confidentiality map traces the joint impact on the disclosure risk R and the data utility U of changes in parameter values of the SDL procedure. In its most basic form, an R-U confidentiality map is the set of paired values (R, U) of disclosure risk and data utility that correspond to various strategies for data release. The map then shows tradeoffs between R and U. It also enables comparison of different SDL procedures. Typically, these strategies implement a disclosure limitation procedure, like masking through the addition of random error. Such procedures are determined by parameters, for instance, the magnitude of the error variance λ^2 for noise addition. As λ^2 is changed, a curve is mapped in the R-U plane. Visually, the R-U confidentiality map portrays the tradeoff between disclosure risk and data utility as λ^2 increases, and so more extensive masking is imposed. Conceptually, the R-U confidentiality map is quite general, and can in principle be applied in any SDL situation. For two different contexts, we illustrate in the next sections how the map can be developed, and also explore some of the context-specific implications of the procedure. Section 6.4.1 deals with SDL through adding multivariate additive noise. It requires familiarity with matrix algebra and statistical theory, and can be skipped by those readers without this background. Section 6.4.2 deals with SDL through topcoding. The mathematical demands of Section 6.4.2 are less than those of Section 6.4.1 and should be accessible to all readers.

6.4.1 Constructing an R-U Confidentiality Map: Multivariate Additive Noise

In this section, we construct an R-U confidentiality map for the SDL technique of multivariate additive noise. For the simpler case of univariate noise, see Brand (2002a), Spruill (1983), Paass (1988), and Duncan and Mukherjee (2000). For an introduction to the multivariate case, see Sullivan and Fuller (1989). Here we follow an implementation of multivariate additive noise that was discussed by Kim (1986), and explored by Kim and Winkler (1995). Our approach is based on a few different, but yet realistic, assumptions about the state of knowledge and motivation of the data snooper. We demonstrate that the assessment of disclosure limitation methods depends critically on the knowledge state of the data snooper.

Although in practice the form of R and U should be tailored to the particular situation at hand, here in order to illustrate the concept of the R-U confidentiality map we take the DSO to have adopted the following two specifications, the first about the data user and the second about the data snooper:

1. *Data user*. The data utility U is the reciprocal of the data user's mean squared error in estimating a given linear combination $\mathbf{c}'\boldsymbol{\mu}$ of the components of the population mean vector $\boldsymbol{\mu}$.
2. *Data snooper*. The disclosure risk R is the reciprocal of the mean squared error the data snooper can achieve in inferring the value τ for a target attribute of an individual entity. Note that this specification means that the DSO is concerned with limiting attribute disclosure, not identity disclosure. With masking through additive noise, even if a data snooper knows which record corresponds to a particular individual, the actual value of an attribute will not be known.

Given these specifications, we now examine two particular states that a data snooper may be in regarding knowledge of the target value τ. First, suppose that the data snooper knows just that τ is like one of the values in the general population. We call this the *Population* knowledge state. In this state the data snooper has the same personal probability distribution for τ as that of the population distribution of the attribute in question. We also consider an alternative knowledge state that we call Record. In this state the data snooper has enough external information to be able to identify the specific record in X to which the target record belongs.[1] The Population knowledge state is appropriately assumed when the data are a small sampling fraction from a population and the data snooper cannot reasonably be sure that the target entity is in the sample. For exposition purposes we will assume that the

[1] Another possibility is the Sample knowledge state where the data snooper is taken to know that τ is one of the values in $\mathbf{X}$ (i.e., is in the sample). The Sample knowledge state is appropriately assumed when the data snooper knows, for whatever reason, that the individual was surveyed in a sample survey. This would certainly be true if the data are a census or a near census. This state is called "response knowledge" by Keller and Bethlehem (1992). Or the data snooper might know that the record of the target is in the sampling frame, but not necessarily in the actual sample.

sample is obtained through simple random sampling. The Record knowledge state is appropriately assumed when the data snooper has sufficient external identifying information to permit linkage to the target record.

As is usual, take the original data to consist of n records, each of p attributes. Express the data in matrix form as $\mathbf{X} = [\mathrm{X}_{ij}] = [\mathbf{X}_1, \ldots, \mathbf{X_n}]'$. Assume that the records are a random sample from a population with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\sum$, so we write $\mathbf{X}_i^{iid} \sim (\boldsymbol{\mu},\ \boldsymbol{\Sigma})$ and the masked data have the additive noise form $\underset{n\times p}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}} + \underset{n\times p}{\boldsymbol{\varepsilon}}$ where $\underset{n\times p}{\boldsymbol{\varepsilon}} \sim (\mathbf{0}, \lambda^2, \Sigma)$.

Data utility: Consistent with common practice, we take the data user to estimate the population mean vector $\boldsymbol{\mu}$ using the ordinary least squares estimator, $\underset{p\times 1}{\hat{\boldsymbol{\mu}}} = \underset{p\times 1}{\bar{\mathbf{Y}}}$, the sample mean vector of the masked data. Therefore, $E(\hat{\boldsymbol{\mu}}) = \boldsymbol{\mu}$ and $Var(\hat{\boldsymbol{\mu}}) = \frac{1+\lambda^2}{n}\boldsymbol{\Sigma}$. For a goal of estimating an arbitrary linear combination $\mathbf{c}'\boldsymbol{\mu}$ take the data utility to be the reciprocal of the mean squared error and so $U = \frac{n}{1+\lambda^2}(\mathbf{c}'\boldsymbol{\Sigma}\mathbf{c})^{-1}$ Note that, appropriately, this data utility measure takes into account the multivariate, correlation structure of the data, whilst the data user is not assumed to know and so make use of $\sum$.

Disclosure risk: Suppose the data snooper's goal is to compromise a specific entity and the data snooper is in the *Population* state of knowledge. With the data snooper having a specific target value τ in attribute j and using $\hat{\tau} = \bar{\mathbf{Y}}_j$, the disclosure risk is

$$R = \frac{1}{E(\tau - \hat{\tau})^2} = \frac{n}{(1+\lambda^2)\sigma_j^2 + n(\mu_j - \tau)^2}. \tag{6.1}$$

Consider now the *Record* state of knowledge, in which the data snooper is able to identify the masked record that corresponds exactly to the target τ. This confronts the DSO with the worst disclosure risk. Here, if the snooper uses $\hat{\tau} = Y_{ij}$, which equals $\tau + \varepsilon_{ij}$, the risk is

$$R = \frac{1}{E(\tau - \hat{\tau})^2} = \frac{1}{E(\varepsilon_{ij})^2} = \frac{1}{\lambda^2\sigma_j^2}. \tag{6.2}$$

Whatever the knowledge state of the data snooper, the disclosure risk without data masking is found by setting $\sigma_j^2 = 0$. Note that under Record state of knowledge, without masking R will be infinite. Thus, in circumstances where record linkage is feasible for the data snooper, release of the original data would obviously pose too much of a threat to confidentiality.

Knowing the target's index i, is the data snooper always better off using $\hat{\tau} = Y_{ij}$ to assess the target value τ? Comparing Equations (6.1) and (6.2), the data snooper actually gains by using $\bar{\mathbf{Y}}_j$ whenever $\lambda^2 > \left(\frac{n}{n-1}\right)\left(\frac{\tau-\mu_j}{\sigma_j}\right)^2 + \frac{1}{n-1}$, which for large n is approximately the square of the number of standard units the target τ lies from the mean. Thus, by adding sufficient noise, the DSO can eliminate the advantage

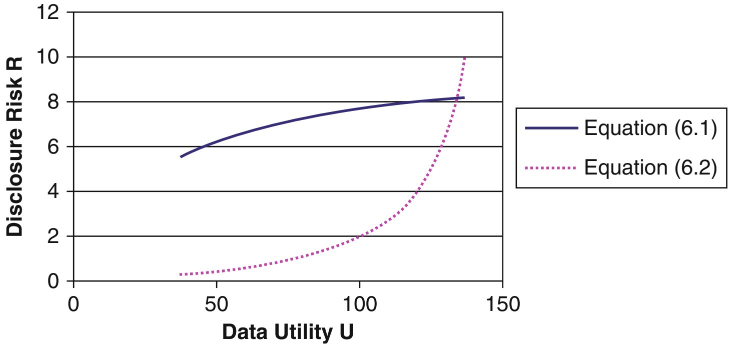


Fig. 6.1 R-U confidentiality map

the data snooper has through record linkage. As the R-U confidentiality map makes visually evident, overcoming this advantage comes with cost in terms of data utility.

Displayed in Fig. 6.1 is an R-U confidentiality map for the two risk measures in this example (with $n = 50$, $\sigma_j^2 = 1$, $(\mathrm{c}'\Sigma\,\mathrm{c})^{-1} = 3$, and $(\mu_\mathrm{j} - \tau)^2 = 0.1$). The map shows the impact on data utility and disclosure risk of changes in the disclosure limitation parameter λ^2 when the data snooper knows the index of the target – Equation (6.2) and when the data snooper does not – Equation (6.1).

Note by examining Fig. 6.1 that with additive noise masking, knowing the index does not help the data snooper once the extent of noise is large enough, specifically when $\lambda^2 > 0.12$, approximately. When the curve for Equation (6.2) crosses below that for Equation (6.1), so $U < 133.6$, approximately, the data snooper is better off ignoring knowledge of the index of the target. If so, the disclosure risk would switch to the upper curve (so for Equation (6.1)) in this domain of larger λ^2.

We have demonstrated the construction of an R-U confidentiality map that allows an assessment of the disclosure risk and data utility under multivariate additive noise. This illustration also allows a comparison of the implications of two different states of data snooper knowledge. Switching SDL technique, in the following section we will construct an R-U confidentiality map for topcoding.

6.4.2 R-U Confidentiality Map for Topcoding

To illustrate topcoding and its impact, consider an example drawn from Duncan and Stokes (2004). For the 1999 New York City Housing and Vacancy Survey (see www.census.gov/hhes/www/housing/nychvs/abstract.html), the Census Bureau determined that 0.5% of the renter-occupied units had a contract rent above $2950. In disseminating characteristics of rental units and their occupants, the Census Bureau, instead of releasing a rent value above a topcode threshold of $\gamma = \$2950$, released the mean rent for these cases (hence mean of all rents conditional on the rent being above \$2950), which is \$3817. To examine the impact on disclosure risk and data utility of this topcoding procedure with $\gamma = \$2950$, *and with other possible choices of* γ, we develop an R-U confidentiality map. As the threshold γ is changed, a curve is mapped in the R-U plane. Graphically, the R-U confidentiality map portrays the tradeoff between disclosure risk and data utility as γ is lowered, and so more extensive masking is imposed. Such a map can aid a DSO in making a decision as to what value of γ to employ.

Suppose that a data analyst wants to examine the relationship between rent and household income. The analyst employs an elementary Cobb–Douglas model, in which expected log rent is taken as a straight-line function of log income. An analyst *naïve* about SDL might ignore the fact that the data have been topcoded, with 81 rent values all replaced with \$3817. Using all 9985 observations of the topcoded data for annual household incomes above \$1000, a regression of log rent on log income yields the fitted regression equation, log rent = 3.320 + 0.3039 log income, with standard error of regression of $S = 0.5287$, standard error of the regression coefficient of log income of 0.005391, and coefficient of determination $R^2 = 24.1\%$.

From a data utility standpoint, we ask what would the results of this regression have been if there had been no disclosure limitation through this topcoding procedure? Without the actual confidential data, this cannot be determined with certainty. However, we can impute the 81 topcoded rent values by simulating them according to the previously fitted regression model, that is, from a normal distribution with mean 3.320 + 0.3039 log income and standard deviation 0.5287. If a simulated value is above 7.99 (corresponding rent \$2950), it is kept, and otherwise another value is simulated. The process is repeated until one simulated value is available for each of the 81 topcoded values. These values are adjusted by adding a constant to give an average imputed value equal to the original data's average, which is 8.25 (so average imputed rent=\$3817).

With this imputation we obtain regression results of log rent = 3.323 + 0.3035 log income, with standard error of the coefficient of log income 0.005392 and coefficient of determination $R^2 = 24.1\%$. The data utility of the topcoded data can be assessed in a variety of ways. Since the estimated coefficient of the regressor is typically of interest, for our illustration let us take data utility to be the reciprocal of the squared difference between the estimated regression coefficient of log income from the original data (as imputed) and the regression coefficient of log income from the topcoded data. With a difference of 0.0004 in the coefficient estimates and dividing by 10^7 for convenience in scaling, this gives 0.625 as the data utility.

Now what about modeling disclosure risk for a particular data product? In general, this can be thought of as target specific. When the concern is identity disclosure, it is the probability of an identification of the target. In any application, this will depend on the circumstances of two key factors: (1) motivation of a data snooper to attack and (2) identifiability of a target. The first factor drives the probability of an attack. The second factor determines the probability of identification given an attack. This can be considered a simplified form of the scenario analysis framework outlined in Chapter 2. There we outline 11 factors effecting a decision over whether to attack or not; motivation being one of those. In the particular case specified here, both of these multiplicative probability factors would reasonably be considered increasing functions of x. To further our illustration, we take for a given released rental value x that the overall probability of identification of a respondent with income x is given by the logit model:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x = -5 + 0.001x$$

This model gives the probability that a data snooper can identify a record when the released rental value is \$500 to be 0.01, and the probability of identification for a released rental value of \$5000 to be 0.5, both plausible values. We take the disclosure risk to be the *maximum* probability of disclosure as given by this logit model. Renters with rental values that have been topcoded are taken to have negligible probability of disclosure. Since the logit model has ρ increasing in x, the disclosure risk is the value of ρ given by the logit model at the topcoding threshold γ. For the threshold used by the Census Bureau of \$2950, the disclosure risk would be 0.11. Thus for this specific implementation, the R-U value is (disclosure risk, data utility) = (0.11, 0.625). In this case, the data utility is high. For example, according to the model fit to the topcoded data, a doubling of income predicts a 23.45% increase in rent, whilst according to the model fit to the imputed data, a doubling predicts a 23.41% increase in rent, hardly a difference.

Since the disclosure risk is not low, we explore what the impact would be on disclosure risk and data utility of lowering the topcoding threshold, say to \$2500 or \$2000. This can be done by constructing an R-U confidentiality map.

To obtain the R-U confidentiality map, we carry out the above process for a variety of values of the topcoding threshold γ. The result for topcoding thresholds ranging from \$2950 to \$1800 is given by the circles in Fig. 6.2. On average we

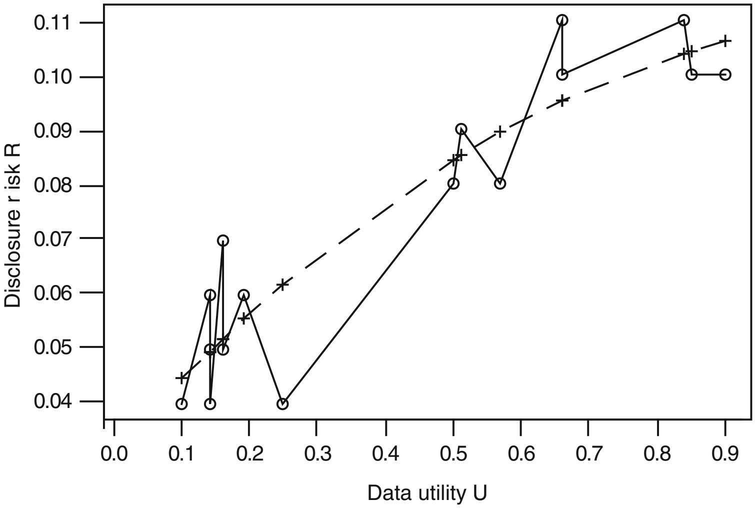


Fig. 6.2 R-U confidentiality map

would expect that increasing γ would lower both the disclosure risk R and the data utility U. This is not the case with specific data values where there can, as in this illustration, be variation from the expected curve. We also display a fitted quadratic curve to smooth the empirical variation. Using the fitted curve we see that to get the disclosure risk below 0.08, say, requires a topcoding threshold value of $\gamma = 2560$. In that case the data utility is about 0.5.

6.5 Discussion

The number of SDL methods has increased significantly in the past few years, but their advantages, compared to one another, have been barely explored. There is an increasing need to design suitable criteria that allow us to compare alternative methods and identify which performs best and when. Usual solutions define suitable measures that can express the extent to which R is low and U is high. This chapter has presented several such measures and their interpretation.

For purposes of implementation, however, we must realize that there is no agreement on what these measures R and U should be in particular applications. Complicating this task is the fact that, disclosure risk and data utility are ambiguous concepts, certainly multivariate in nature and that utility as measured typically is only a proxy for the real underlying utility, which the DSO will not know before the data product is released. Further, for a given disclosure limitation problem, the DSO may identify several classes of potential data snoopers, for each snooper several data targets and for a given data snooper and a given target several alternative forms of snooper attack.

Masked data from different SDL methods have correspondingly different characteristics. Think of our example of masking income data by coarsening. Consider comparing this to adding noise to the income values. In relating income to education, with noise addition the regression is uncomplicated and the interpretation of the results straightforward. This is not true with coarsening, where the coarsening would require a change of regression technique. These differences make it difficult to have a single measure of data utility and disclosure risk with which to compare SDL methods. Even assuming there was one method, it is not obvious how to define R and U to compare different masked data. Further, even with agreement on R and U, there can be no linear ordering of (R, U) pairs. The problem is made even more complex by the fact that the values of these (R, U) measures may be uncertain from the DSO's perspective. In fact, R and U for the released data depend on the both the data snooper's future actions and the data users' future actions. These can at best be partially known to the DSO, which must be uncertain about the data snooper's motivations and capabilities and the data users' prior information, the inference problems they will address, the estimation procedures that they use, and so on. As Mackey and Elliot (2009) point out complex game theory models may be necessary to obtain some sort of leverage on this. However, notwithstanding these complications, this chapter has laid out a foundation for the ongoing discussions of how best to use SDL methods in dealing with the tradeoff between disclosure risk and data utility.

There are several further avenues for development of the R-U confidentiality map that would be useful. First, we find value in developing analytical R-U maps for some more complex SDL methods, such as data swapping (Dalenius and Reiss, 1982) and the generation of synthetic or virtual data (Rubin, 1993; Abowd and Woodcock, 2004). Second, the R-U confidentiality map itself could be generalized to address more complex decisions about a SDL choice. For example, it might be useful to combine R-U maps for those cases where the utility of many different parameter estimates must be considered, through plotting a weighted average or maximum of a small set of R and U values. Finally, an interesting adaptation of the concept of an R-U confidentiality map would be one that allows exploration of the risk and utility tradeoff for an SDL procedure indexed by two or more parameters, which is suggested by disclosure limitation through microaggregation and binning (see Domingo-Ferrer and Torra (2001)).

Chapter 7
Restrictions on Data Access

Thus far, our focus has been on achieving confidentiality protection through SDL procedures, which as we know mask the data that are to be disseminated. This is the *restricted data* approach.[1] As we noted in Chapters 1 and 2, a different but complementary approach is for the DSO to instead control the process of accessing and analyzing the data.[2] Restricted access has the potential advantage for users who can subscribe to the DSO's stipulations that they will often be able to obtain data that are richer than SDL masked data, typically with more geographical and temporal detail.[3,4]

DSOs have always restricted access to their data, often without the alternative of a masked version, by enforcing certain conditions of data access. For example, DSOs have specified who can have access to the data, at what locations and for what purposes.[5,6] As noted in National Academy Press (2005), "The actual data collected for statistical purposes from households, individuals, business establishments, and other organizations through censuses and surveys under a pledge of confidentiality are never made available to users. Instead, data are made available either in the form of confidential, restricted-access data files or in the form of anonymized data products, including published tables and microdata files."[7]

To some extent, very severe access restrictions will always be with us for some types of data; for example, even the most libertarian of data users would not expect to have access to military intelligence services data – masked or not! (But note the WikiLeaks phenomenon – wikileaks.ch!). More generally, heavily access-restricted data are still the norm for much of what is called administrative data (data collected

[1] Instead of the terms *restricted data* and *restricted access*, Marsh et al. (1991) proposes the terms *safe data* and *safe setting*. We prefer the former as the latter could be taken to imply 100% or even precisely calculable guarantees which would be unrealistic.

[2] www.fcsm.gov/committees/cdac/cdacpaper.pdf

[3] aspe.hhs.gov/hsp/leavers99/datafiles/ch_5.pdf

[4] iga.ucdavis.edu/LDA/restricted-access-data

[5] www.census.gov/prod/2003pubs/conmono2.pdf

[6] www.fcsm.gov/working-papers/SPWP22_rev_ch1.pdf

[7] www.nap.edu/openbook.php?record_id=11434&page=66

G.T. Duncan et al., *Statistical Confidentiality*, Statistics for Social
and Behavioral Sciences, DOI 10.1007/978-1-4419-7802-8_7,

for administrative purposes and not for research purposes), with typically only (some) employees of the DSO having access. However, research access to administrative data is now being opened up (through, for example, initiatives such as the United Kingdom's Administrative Data Liaison Service[8]).

To give an example, in 1982 the US Census Bureau established the Center for Economic Studies (CES)[9] with one of its purposes being to broaden access to the agency's economic microdata for the manufacturing sector. Access is not freely available; indeed CES policy has been to restrict access according to several conditions. As a start, researchers must submit proposals to the CES that specify the research they want to conduct, the economic microdata they will need, and the nature of the results they expect to publish. Only if the proposed research is deemed beneficial to the Census Bureau's programs can access be approved by the agency. To actually get access to confidential data, the researchers must obtain Special Sworn Status taking the Title 13 oath of non-disclosure just as Census Bureau employees must do. Additional requirements must be met to access tax return microdata.[10]

Broadly the restricted access approach can be divided into four components – the "who," "where," "what," and "how" controls:

- *Who*—who has access to a given data set?
- *Where*—where is the data access/analysis to be carried out?
- *What*—what analyses may or may not be conducted?
- *How*—how is access obtained?

These controls are interrelated—a decision about one has implications for the others. We will consider each of these controls in turn.

7.1 Who Can Have Access?

The essence of the "who" question is how the DSO identifies "safe people" or "trusted researchers." It establishes those individuals or organizations that the DSO trusts. Individuals who have a track record in good practice of data security and stewardship may be given greater data access rights than those who do not. This moves away from the straightforward notion of public-use data files.

Often, restricted access conditions stipulate that researchers must have specified credentials to get access to data. In the CES example above, only researchers in economic fields may have access. As another example, IPUMS International provides

[8] http://www.adls.ac.uk

[9] http://www.ces.census.gov/

[10] http://www.census.gov/prod/2003pubs/conmono2.pdf

access exclusively to “bona-fide users” who affirm to abide by a non-disclosure agreement.[11]

In general, what are the qualifications demanded of researchers before they can obtain access to confidential data? Here are some typical qualifications by which a DSO judges a potential data user:

1. Capability of doing research of scientific merit that is intended to be published. The Statistics Canada Research Data Centres Program, for example, requires that an applicant “Include any identifiable contributions made by the applicants to the advancement, development and transmission of knowledge related to the disciplines supported by SSHRC. This element will help assess if the research team members have the expertise and ability to carry out the work.”[12]
2. Proposed research can be accomplished using the data sought, and cannot be accomplished with public-use data files.
3. User must be associated with some institution that can assure compliance with DSO data access requirements. With the Statistics Canada Research Data Centres Program, for example, a researcher from a post-secondary institution outside their network has to have the approval of the academic director of the RDC before the application can even be considered.
4. Proposed research must benefit programs of the DSO.
5. Pass some form of security screening. For the RDCs of the US Census Bureau this requires obtaining Special Sworn Status and so researchers must undergo a security check, including fingerprinting. For the Statistics Canada Research Data Centres Program applicants must pass an “Enhanced Reliability Check” and take “The Oath or Affirmation of Office and Secrecy.”

Note that point 4 suggests a “Why” criterion.

7.2 Where Can Access Be Obtained?

In order to better control how data users may interact with the stored data, DSOs may require that users access the data only at sites that are under its jurisdiction. Generally DSOs are more comfortable in providing access where they can oversee the activity of data users. On the other hand, data users have several complaints about such an arrangement: (1) travel to one of these sites is expensive, (2) the facility is only open at certain hours, (3) computing facilities may be unfamiliar or inadequate, and (4) Internet access may not be available.

Continuing with the CES example, because of the expense of travel, the Census Bureau in 1994 established a regional Research Data Center (RDC) at their Boston office. With the support of Carnegie Mellon University, in 1997 they opened their first academic RDC. Helped by National Science Foundation support, the

[11] international.ipums.org/international/

[12] www.statcan.gc.ca/rdc-cdr/index-eng.htm

Census Bureau followed with RDCs at universities in California, Research Triangle Park, Chicago, and Michigan. More recently, research data centers (RDCs) have been established by several agencies. In the United States, they include a network of RDCs established by the US Census Bureau,[13] National Center for Health Statistics,[14] and the Agency for Health Care Research and Quality.[15] Other countries also maintain research data centers. In Germany, for example, there is an RDC at the Institute of Economic Research,[16] an RDC at the Institute for Educational Progress, and an RDC at Humboldt University in Berlin. SHARE (Survey of Health, Aging, and Retirement in Europe) data are distributed through the RDC at Tilburg University in the Netherlands. The Research Data Centres Program of Statistics Canada is a network of 15 Research Data Centers, 6 branch RDCs, and the Federal Research Data Centre in Ottawa. Three of these RDCs are the Atlantic Research Centre at Dalhousie University, Toronto Regional Statistics Canada RDC at the University of Toronto, and the British Columbia Inter-University Research Data Centre located at the University of British Columbia.

7.3 What Analysis Is Permitted?

In the CES example only those who are able to produce a research proposal that is deemed to be beneficial to Census Bureau programs may have access. This requirement is motivated by the legal basis for Census Bureau operations. It is not a widespread requirement. More typically, data users must promise that they will not carry out analyses that would result in disclosures. Also the intended societal benefits are broader. For example, with the Statistics Canada Research Data Centres Program the stated benefits are to "provide opportunities to

1. generate a wide perspective on Canada's social landscape;
2. provide social science research facilities across the country in both larger and smaller population centers;
3. expand the collaboration between Statistics Canada, SSHRC [Social Sciences and Humanities Research Council], CIHR [Canadian Institutes of Health Research], CFI [Canada Foundation for Innovation], universities, and academic researchers and build on the Data Liberation Initiative; and
4. train a new generation of Canadian quantitative social scientists.[17]

[13] see, www.ces.census.gov/index.php/ces/researchprogram and also e.g., www.ccrdc.ucla.edu/ and http://econ.duke.edu/tcrdc/,

[14] www.cdc.gov/nchs/r&d/rdc.htm

[15] Restricted access as used by US Federal statistical agencies is described in www.amstat.org/sections/srms/proceedings/y2004/files/Jsm2004-000581.pdf

[16] www.diw.de/english/soep/research_data_center_of_soep/27903.html

[17] http://www.statcan.gc.ca/rdc-cdr/index-eng.htm

7.4 Modes of Access

There are four modes of access that are currently used by DSOs for disseminating data for use outside of their organizational boundaries:

1. Free access
2. Delivered access
3. Safe settings
4. Virtual access

7.4.1 Free Access

Free access (or what can be called *unrestricted access*) is used for publishing some census tabulations and headline administrative data. An instance of free access is the United Kingdom's Neighbourhood Statistics Service.[18] Neighbourhood Statistics is intended for public-use data. It imposes no restrictions on who can access the data nor on what they can do with the data. Also, there is usually no monitoring of who is accessing this data and what they are doing with it. Until 20 years ago, the dissemination medium of such data was paper-based, usually in the form of thick volumes of tables. However, web delivery is now far more common, and this has arguably opened up data sets for much wider use.

7.4.2 Delivered Access

Delivered access is a more restricted form of access, in which access to the data is applied for and the data are delivered to the user through some physical medium such as CDs or now more commonly through an Internet portal. Here is a description of an example of delivered access: "The National Center for Education Research (NCER) has released a CD-ROM that contains documentation, an electronic codebook, and restricted-use data files for the fall 2002 and spring 2003 waves of data collection for the Preschool Curriculum Evaluation Research Pilot Year Study. Data from direct child assessments, parent interviews, teacher report on child behavior, preschool teacher interviews, and classroom observations are provided in ASCII format. These CD-ROMs are only available to holders of restricted-use data licenses."[19,20]

[18] The *UK's Neighbourhood Statistics Service* (www.data4nr.net) provides local area indicators derived from administrative records of multiple government agencies.

[19] Other examples include the UK Data Archive (www.data-archive.ac.uk) and the Integrated Public Use Microdata Service (IPUMS) (international.ipums.org/international/).

[20] ies.ed.gov/ncer/pubs/

Usually, as in this NCER case, the process of applying for a copy of the data requires the user to specify what it is to be used for. Also, the user is required to agree to specified conditions as a license for data access. The range of data sets that are covered by delivered access is wide, ranging from the public-use microdata files (PUMFs) provided by some DSOs with minimal access restrictions and with broad potential-user bases to disclosive data sets which might, for example, be given only to a few trusted researchers under highly restrictive license conditions.

7.4.3 Safe Settings

Safe settings are regarded as the strongest form of restricted access. Here the data user applies for access to use the data in a particular location—often in the offices of the DSO, otherwise at a research data center (RDC) that has been established by the DSO. Usually, the user is required to analyze the data on a dedicated, stand-alone computer. The user is not permitted to take in mobile data transport devices such as CDs or USB flash drives. The user is allowed to take away certain analytical output, but only after it has been checked by the DSO for disclosiveness.

The safe settings level of "where" control is generally regarded as undesirable by researchers because it requires users to travel to a DSO-specified location and to work in unfamiliar surroundings. Under safe settings, therefore, utilization is often low. An alternative approach (used for example by the UK longitudinal study) is that the researcher submits software syntax for their analysis to a dedicated unit which if it approves then runs it. However, this rather awkward approach is being superseded by virtual access systems.

7.4.4 Virtual Access

Virtual access is now widely regarded as the future of research data access. It combines many of the advantages of the physical safe setting with much of the flexibility of having a copy of the data on one's desktop. There are two variants on the virtual access theme: *Direct access* and *analysis servers*.

Direct virtual access uses virtual remote network interfaces to allow users to view, interrogate, manipulate and analyze the data as if it was on their own machine. There are two critical differences between direct virtual access and delivered access. First, typically with virtual access output is checked by a member of staff at the DSO; this enables the DSO to potentially spot data abuses.[21] Second, there is no possibility of a user/snooper directly linking the accessed data set to a source of identification information, and this greatly restricts the type of disclosure scenarios that the DSO needs to consider.

[21] Although service demands to check and return output promptly do reduce the value of this somewhat.

Analysis servers go one step further in not allowing direct access to a data set, but rather allowing the user to interrogate it. In such systems, data can be analyzed but not viewed. Usually, there is a mechanism for delivering the analysis (for example, through uploading scripts for common statistical packages or, occasionally, though a bespoke interface). The analysis server will return the output from the resultant analysis—usually again after it have been checked by DSO staff for disclosiveness. From the DSO's view point, the advantage over direct virtual access is that because the user cannot see the data the risk of spontaneous recognition of a data unit is all but removed. The disadvantage from the user's point of view is that it is now more difficult to explore the data.

We have talked about the checking of output for "disclosiveness" without considering what this term means. Our starting assumption in developing a definition of "disclosiveness" is that the data are regarded as too risky for delivered access to be used. Therefore, what is being checked is whether it would be possible to recover (some of) the underlying data from the output. As a simple example, with risky tabular frequency data a DSO typically does not permit unrestricted requests of multivariate tables of counts, since a sequence of such requests can be used to recover the original data (see, e.g., Duncan et al., 1999). So, if a user were to request such cross-tabulations, then the request would have to be denied. However, the problem goes beyond this situation. For example, using any regression model in combination with a set of residual plots, it is possible to recover some of the original data used to generate the model. At this point, developing valid output checking processes that are automated is an open research question. Therefore, output needs to be checked manually and such checking needs to be carried out by DSO staff with some expertise. One way to reduce the burden on the output checkers, used by, for example, ONS's virtual microdata laboratory,[22] is to define a restricted class of outputs as "safe" and then leave it to the user to demonstrate that anything that does not fall on that list is also safe.

7.4.5 Licensing

Another tool a DSO has is a licensing mechanism, often used in conjunction with other restricted access mechanisms. For example, the US National Center for Education Statistics licenses researchers to use, at their university or research center, data sets that contain richer data than the public-use file. Some common themes in such licensing agreements are as follows:

1. Specification of those permitted access
2. Data security and enforcement/provision for inspection
3. Restrictions on use (particularly, any prohibition against linking with other files)
4. Policy on return or destruction of data provided

[22] www.ons.gov.uk/about/who-we-are/our-services/vml

The function of licensing is threefold:

1. Consistent with the “who” question of Section 7.1, it clearly distinguishes between those individuals or organizations the DSO trusts and those that it does not.
2. Consistent with the “where,” “what,” and “how” questions of Sections 7.2, 7.3 and 7.4, it is a framework for specifying the conditions under which access can occur.
3. It can specify sanctions or penalties should the individual/organization transgress on those access conditions.

It is possible to have licensing at graded levels, with different users having access to data with different levels of disclosure risk (and therefore presumably different levels of data utility). In the United Kingdom, the Office of National Statistics (ONS) makes a distinction between public, research, and special license levels of access. A problem with this approach is that it is open to abuse by the DSO (who might choose to deny access to a researcher that they do not like or whose proposed research they do not like). An inexperienced researcher may be subject to harsher conditions than a professor of long standing. So, as a general mechanism for the dissemination of data for research purposes it can be criticized on fairness grounds.

Function 2 above concerns both the “how” and “where” of access. Obviously by laying out the conditions of access the DSO maintains some control over the security of the data. Also, in effect the DSO can provide guidance to the data user over good practice. If the data are being provided to the user on site, then various physical and computer security conditions might be required. Here are some examples of possible requirements:

1. Data are stored in dedicated secure data lab.
2. An independent locking system (unmastered) to the data storage area is used.
3. Extra security at all possible primary and secondary points of entry. Extra locks on doors, bars on windows, etc.
4. Data are stored on a stand-alone machine with multiple passwords required to access the data.
5. Devices such as external disk drives/USB ports are disabled.
6. Output is not removed from the secure data lab and is destroyed when finished with.
7. Entry to the data lab is limited to particular staff.
8. Log books of access are kept.

As well as providing actual security, imposing such conditions may also be intended to change the mindset of the user, who will hopefully react to them by being more security aware. The flip side of this is that these conditions may place awkward obstacles in the researcher’s usual research path.

Another type of license condition which is commonly employed asks the user to agree to restrictions on what they can do with the data—in particular, not linking it with other data sets that contain identifiers.

Function 3 of the licensing process involves the sanctions that can be applied to users or their organizations for non-compliance with the license conditions. In order to serve as deterrents for non-compliance, they must be enforceable. Typical sanctions are fines and removal of the right to access the data. For example, with the Statistics Canada Research Data Centres Program: “Researchers whose projects are approved will be subject to a security check before being sworn in under the Statistics Act as ‘deemed employees.’ Deemed employees are subject to all the conditions and penalties of regular Statistics Canada employees, including fines and/or imprisonment for breach of confidentiality.”[23]

The threat of sanctions will be taken most seriously if the data user or their organization is subject to a security audit by the DSO. While an audit can be costly to both the DSO and the user, a license without such a stipulation may not be taken seriously.

Overall, licensing can be a useful way to decrease disclosure risks for certain uses of a disclosive data set by researchers especially when explicit or implicit sanctions can be invoked. Although it is commonplace for users to have to sign access agreements for routine data access requirements, beyond giving the user cause for reflection at the point of access there is little or no enforceability in such agreements. The question of whether agreements that are not directly enforceable have real impact is of course a major question in many areas of society. Answers will vary depending on history, social context, existence of informal controls, etc.

7.5 Conclusion

The primary advantage of restricted access methods is that it can provide qualified researchers with rich microdata. The primary disadvantage of these methods is the burden that it places on researchers. As noted by McCaa and Ruggles (2002), “The problem with data enclaves such as the U.S. Census Bureau Research Data Centers (RDCS) is that they impose heavy costs on social science and policy research. It is not easy to use a Census Bureau Research Data Center. Because of the cost barriers and inconvenience, the RDCS have attracted few researchers. Only well-funded investigators doing work deemed valuable by the Bureau are eligible to use the centers. The user registration logs for the IPUMS data extraction system[24] suggest that a majority of microdata users are graduate students, who would for practical purposes be excluded from using an RDC. [Note though that Statistics Canada does have special provisions for graduate students as well as undergraduate students as co-applicants with their advisors in its Research Data Centres Program.] Even if the funding problem could somehow be overcome, the number of seats in the centers would have to be multiplied a thousand fold to accommodate the current number of users of public microdata. The RDCS were never intended as a substitute for public use microdata and they cannot fulfill that role.”

[23] www.statcan.gc.ca/rdc-cdr/faq-eng.htm#q8

[24] www.ipums.org

On the other hand, the restricted data methods used to produce public microdata place no such burdens, but the quality of the data may not be adequate for many research purposes and the SDL process places a burden on the DSO. As McCaa and Ruggles (2002) state, “The alternative to restricted access is data modification. The most straightforward data modification is the reduction of detail, but researchers have expressed alarm at this alternative.” In May 2000, Ruggles was asked by the Census Bureau to report on the potential impact on users of reducing the detail offered by the 2000 Public Use Microdata Sample (PUMS) of the US census, which was then being contemplated as a means of reducing the risk of disclosure. Accordingly, approximately 1300 current users of the IPUMS-USA data were e-mailed and requested that they fill out a web-based survey on the issue. Within 7 days, 1006 researchers had completed the survey. The reaction was remarkably uniform: data users overwhelmingly expressed a preference for maximum detail and described hundreds of research projects that would have to be abandoned if the Bureau reduced detail significantly. Many users were outraged by the suggestion that subject detail might be reduced; one wrote, for example, “As far as I am concerned, elimination of the detail of age, race, ancestry, income, occupation, and geography would essentially eliminate the value of data from the long form. This is a shameful, cowardly, and ludicrous proposal. I hope it will disappear promptly and not be raised again” (Ruggles, 2000). Of course users also complain about access restrictions and therefore as with most of the topics covered in this book the DSO has to tradeoff between apparently negative and certainly difficult choices.

Chapter 8
Thoughts on the Future

> *We are living in exponential times. We are currently preparing students for jobs that don′t yet exist using technologies that have not yet been invented in order to solve problems we don′t even know are problems.*
>
> —Karl Fisch and Scott McLoed, 2007[1]

The future will surely bring challenges to statistical confidentiality. Some challenges will be familiar, much like the ones described in Chapter 1. But as the lead quotation suggests, we must prepare for exponential change in our responsibilities, the technology we employ, and the problems we face. Specifically, we must prepare for dramatic changes both in information technology and in our social, economic and political environment. This chapter lays out our view of how these changes will multiply the tensions between the demand for the protections of confidentiality and the demand for access to data. Interestingly, much of what we discuss was only hinted at two decades ago by Duncan and Pearson (1991). Their glimpse to the future is today's reality. Anticipating tomorrow requires a new and expanded forecast.

Our forecast for the next decades is one of accelerating technical capacities for capturing, storing, integrating, and disseminating personal data. In Chapter 1 we labeled this data process CSID for *C*apture, *S*tore, *I*ntegrate and *D*isseminate. This acceleration is consistent with the recent experience of lower costs for surveillance, computer storage, linking of databases and speeding information about the world. It opens opportunities for Data Stewardship Organizations (DSOs) to do more with personal data. Professional staff in DSOs, facing fewer constraints on what they can do, will be under pressure to fully respond to new demands for data access.

Demands for personal data will further accelerate as researchers and policy analysts seek to provide answers to new questions about a changing environment. Sieber (2007) notes, "As social and behavioral research expands to involve diverse populations, contexts, and sensitive topics, it raises many complex issues of privacy and confidentiality. These issues go beyond what is contained in textbooks or known to most researchers and Institutional Review Board (IRB) members."

[1] Did You Know 2.0 www.youtube.com/watch?v=pMcfrLYDm2U

G.T. Duncan et al., *Statistical Confidentiality*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-7802-8_8,

Consider this topical research question: “What do we understand by a *community*?” The links that form communities today have morphed beyond simple geographical characterization—being a “neighbor” no longer need suggest that we live next door. Instead, Wi-Fi broadband access and mobile devices have allowed virtual social communities to form. In the 1990s, they would be said to exist in cyberspace and be treated as a world apart from our normal existence. Now this phenomenon is so much part of accepted reality and is so entwined with our daily existence that it hardly deserves such a conceptual separation. As Elliot (2011) argued, our online activity is as much a part of our identities as our offline activities. Indeed to give a specific illustration, even matrimonial ties within certain dispersed groups are now almost routinely facilitated by websites such as Indian Friend Finder.[2] Even more remarkably, by October 2009 more than 16 million people spent many of their hours and their dollars in the artificial world of *Second Life*,[3] which is a 3D digital world imagined and created by its residents. At a specific moment on September 16, 2007, the *Second Life* website said 46,354 people were currently online and that US $1,099,947 had been spent in the previous 24 h. In March 2009, it has become known that there exist a few *Second Life* entrepreneurs whose profits exceed US $1 million per year.[4]

Although debates abound about how communities evolve and function, real understanding comes from data. So, statistical information about individual people has become ever more important. How do social controls work? Do hierarchies form? How is conflict managed? How are resources allocated? Data needed to answer such questions are personal. Access to such data must respect statistical confidentiality.

In general, the future environment for statistical confidentiality will be demanding, requiring responses to a range of questions that we address in the coming sections:

Section 8.1. Will privacy and statistical confidentiality have new meanings?
Section 8.2. Who will care about statistical data?
Section 8.3. What new forms of DSOs will develop?
Section 8.4. Will statistical data remain valuable?
Section 8.5. What new data types will need to be addressed?
Section 8.6. What are the implications of new research in computer science about privacy?
Section 8.7. Will there be other new issues for statistical confidentiality?
Section 8.8. Will there be new forms of data snooping?
Section 8.9. What new strategies of disclosure limitation should be developed?

[2] indianfriendfinder.com/

[3] en.wikipedia.org/wiki/Second_Life

[4] Wagner J. Au (March 24, 2009). “Top Second Life Entrepreneur Cashing Out US$1.7 Million Yearly; Furnishing, Events Management Among Top Earners”. nwn.blogs.com/nwn/2009/03/million.html

Let us begin with the question of whether we will have to recast what we mean by statistical confidentiality.

8.1 New Meanings for Privacy and Statistical Confidentiality

Chapter 1 carefully laid out meanings for privacy and statistical confidentiality which are currently workable for DSOs. We expect that in the future, and particularly in some contexts, two of the conceptual distinctions that we made will begin to blur. These two distinctions are between (1) privacy and confidentiality and (2) identity disclosure and attribute disclosure.

Regarding distinction (1), conceptual clarity about privacy and confidentiality issues has been important in developing appropriate policies for DSOs. Privacy issues for DSOs relate to the intrusiveness of data capture. Confidentiality issues relate to data dissemination.

It is well-said that "Privacy means different things to different people, including the scholars who study it, and raises different concerns at different levels" (Acquisti, 2004). Privacy is a personal construct that varies from person to person. It is also a sociocultural construct that will change with broad social changes such as globalization and the growth of the information society. But a common thread among the various conceptualizations is that privacy respects a person's control on information about themselves (Boruch and Cecil, 1979). Accordingly, autonomy rather than the commonplace notion of secrecy is *the* issue for privacy. Confidentiality, on the other hand, refers to the DSO's policy about how identifiable personal information will be managed and disseminated. Broadly, we can say that privacy concerns people, whereas confidentiality concerns data. Confidentiality only becomes operative once personal data are obtained.

In *The Transparent Society*, Brin (1998) makes the case (also made by Dan Klein[5]) that much of the concern about privacy turns out to be a concern about confidentiality: "In so many cases where [privacy] is used the issue would be more aptly discussed as one of confidentiality in transactions or in information shared in completing transactions" (p. 16).[6] The importance of this distinction in forming policy is that confidentiality requires protection of personal data rather than control of data gathering.

We anticipate that computer technology will bring the meaning of privacy and confidentiality closer together. In some cases, software may permit data that are gathered today to be disseminated almost instantaneously. This makes the acts of data collection and data dissemination hard to separate. At the same time, computer technology and particularly new sociotechnical movements such as Web 2.0[7] open

[5] www.fee.org/publications/the-freeman/article.asp?aid=1804

[6] www.ncpa.org/debate2/transpar.html

[7] en.wikipedia.org/wiki/Web_2.0

up the possibility of more individual-level control over how data about them are used, shared and disseminated.

Perhaps more importantly, if individuals have little confidence, whether through evidence or prejudice, in the ability of a DSO to maintain confidentiality of their personal record, they will be increasingly reluctant to provide information about themselves. However, suggestions from the US 2000 Census are that many people thought that providing information was a violation of their privacy.[8] As Ken Prewitt, a former Director of the US Census Bureau, summarized their view, "I don't care if it isn't shared, I still don't want anyone to know that."[9] As privacy sensitivity grows—which we expect that it might well—we can expect that this concern will also grow.

Another trend that may blur the distinction between privacy and confidentiality is the move to obtain more statistical information from administrative records rather than from survey or census activities. In a survey, participants can be promised confidentiality in the statistical use of the data. In capturing administrative data, individual privacy is a concern and participants typically receive assurances about it. To then have to also provide assurances about confidentiality in statistical uses of the data, must be confusing to the participant. However, it could be argued, that the reuse of administrative data for statistical purposes has the potential to increase individual privacy. If such data are being used in the place of yet another survey, then the individual will have taken part in one less data collection exercise and his or her personal details are on one fewer database.

Regarding distinction (2), our understandings of disclosure risk may also change with implementation of new technologies. For example, with video surveillance there is a blurring of identity disclosure (I know that person!) and attribute disclosure (I see what that person is up to!). The camera gives us both simultaneously.

A final point here concerns the rapid expansion of cyber activity. As individuals spend more of their time interacting with others online, their identities and attributes become increasingly present on the net, controlled by themselves on the net, and indeed changed by the net. From simple retail and banking transactions through social networking sites such as *Facebook* through complex highly interactive multi-player games such as *World of Warcraft*[10] or *Travian*[11] to full-fledged cyber lives (for example *Second Life*[12]), our cyberactivity is becoming and increasingly important component of our daily lives. In this context, the distinctions between identity and attribute, privacy and confidentiality, data and person become increasingly blurred. Given this, it is now fair to say that our personal identities are in part cyber-identities. As the immersive power of technology increases, this trend can only continue and this in turn is likely to have further impacts on the meanings of confidentiality and privacy.

[8] www.census.gov/srd/papers/pdf/rsm2002-01.pdf

[9] www.prb.org/Articles/2001/FormerUSCensusBureauDirectorKenPrewittPondersCensus2000.aspx

[10] www.worldofwarcraft.com/index.xml

[11] www.travian.com/

[12] secondlife.com

8.2 Who Will Care About Statistical Data?

The likely expansion of globalization coupled with lowered costs of data processing makes possible new uses of statistical data for wider and more numerous collections of data users:

- With inexpensive web access, there will be more demand for statistical data from students, analysts from smaller agencies at all levels of government, and researchers for a greater variety of non-governmental organizations. There will be increased demand for personal data as they become available in these relatively new forms:

  1. Geospatial
  2. Biometric recognition
  3. Audio and video
  4. Biological material
  5. Electronic network

  We discuss each of these further in Section 8.5.2.

- International data will be increasingly sought by both governmental and non-governmental bodies.

Some of the implications of developments on the international front are evident in the Istanbul Declaration at the close of the second OECD World Forum on Statistics, Knowledge and Policy on June 30, 2007:

ISTANBUL DECLARATION[13]

> We, the representatives of the European Commission, the Organization for Economic Cooperation and Development, the Organization of the Islamic Conference, the United Nations, the United Nations Development Programme and the World Bank, recognize that while our societies have become more complex, they are more closely linked than ever. Yet they retain differences in history, culture, and in economic and social development.
>
> We are encouraged that initiatives to measure societal progress through statistical indicators have been launched in several countries and on all continents. Although these initiatives are based on different methodologies, cultural and intellectual paradigms, and degrees of involvement of key stakeholders, they reveal an emerging consensus on the need to undertake the measurement of societal progress in every country, going beyond conventional economic measures such as GDP per capita. Indeed, the United Nation's system of indicators to measure progress towards the Millennium Development Goals (MDGs) is a step in that direction.
>
> A culture of evidence-based decision making has to be promoted at all levels, to increase the welfare of societies. And in the "information age," welfare depends in part on transparent and accountable public policy making. The availability of statistical indicators of economic, social, and environmental outcomes and their dissemination to citizens can contribute to promoting good governance and the improvement of democratic processes. It can strengthen

[13] http://www.oecd.org/dataoecd/14/46/38883774.pdf

> citizens' capacity to influence the goals of the societies they live in through debate and consensus building, and increase the accountability of public policies.
>
> We affirm our commitment to measuring and fostering the progress of societies in all their dimensions and to supporting initiatives at the country level. We urge statistical offices, public and private organizations, and academic experts to work alongside representatives of their communities to produce high-quality, facts-based information that can be used by all of society to form a shared view of societal well-being and its evolution over time.
>
> Official statistics are a key "public good" that foster the progress of societies. The development of indicators of societal progress offers an opportunity to reinforce the role of national statistical authorities as key providers of relevant, reliable, timely and comparable data and the indicators required for national and international reporting. We encourage governments to invest resources to develop reliable data and indicators according to the "Fundamental Principles of Official Statistics" adopted by the United Nations in 1994.
>
> To take this work forward we need to
>
> - encourage communities to consider for themselves what "progress" means in the 21st century
> - share best practices on the measurement of societal progress and increase the awareness of the need to do so using sound and reliable methodologies
> - stimulate international debate, based on solid statistical data and indicators, on both global issues of societal progress and comparisons of such progress
> - produce a broader, shared, public understanding of changing conditions, whilst highlighting areas of significant change or inadequate knowledge
> - advocate appropriate investment in building statistical capacity, especially in developing countries, to improve the availability of data and indicators needed to guide development programs and report on progress toward international goals, such as the Millennium Development Goals
>
> Much work remains to be done, and the commitment of all partners is essential if we are to meet the demand that is emerging from our societies. We recognize that efforts will be commensurate with the capacity of countries at different levels of development. We invite both public and private organizations to contribute to this ambitious effort to foster the world's progress and we welcome initiatives at the local, regional, national and international levels.

This Istanbul Declaration makes a compelling case for how having high quality statistical indicators helps achieve societal goals. Because these indicators are constructs based on microdata obtained from individuals and organizations, implementation of the Istanbul Declaration requires careful attention to statistical confidentiality issues.

8.3 What New Forms of Data Stewardship Organizations Will Develop?

As organizations progressively value data as a strategic asset, they are compelled by natural ethics, legal enforcement and sociocultural pressures to take on data stewardship as an essential function. For example, the Royal Statistical Society and the UK Data Archive have included explicit principles of data stewardship in its RSS Code of Best Practice.[14] To fulfill the responsibilities that this implies, organizations are increasingly setting up infrastructure to assess and manage their data environments. This may, for example, involve the formation of data stewardship committees and councils to develop policies for knowledge management, and procedures for data flows to, within, and from the organization. Depending on the nature of the organization and its data environment, data stewardship will place different emphases in protecting data assets on confidentiality, data quality and accessibility.

Here are some illustrations of organizational devices DSOs have employed in acknowledging their data stewardship responsibilities regarding confidentiality and data sharing:

1. *U.S. Census Bureau Data Stewardship Executive Policy Committee.* Serves as the focal point for decision making and communication on policy issues related to privacy, security, confidentiality, and administrative records.[15]
2. *University of California at Berkeley Data Stewardship Council.* Deals with issues of data access, privacy, security and use, data governance, data sharing, data integration, data warehousing, information architecture, data quality, data standards and metadata management.[16]
3. *The International Council for Science (ICSU).* Operates 49 World Data Centers globally to form a data network[17] engaged in
    a. building consistent and high-quality records of environmental observations with associated metadata;
    b. partnering with the scientific community (and others) through provision of high-quality data and services for use in scientific studies;
    c. use of new and innovative technology to provide access to data and products.
4. *Direct Marketing Association Data Stewardship Council.* Educates marketers on the responsible use of data and develops best practices and other tools that will help marketers better manage consumer information.[18]

These developments are a signal that organizations will have to put into place systematic administrative procedures for how they deal with personal data. An important component of this process can be summarized as "Privacy by Design" (Duncan, 2007). Achieving an acceptable level of privacy will require that DSOs adopt new information technologies and commit themselves to a high level of managerial attention.

[14] www.data-archive.ac.uk/news/publications/PreservingSharing.pdf

[15] www.census.gov/privacy/files/related_information/003352.html

[16] datasteward.berkeley.edu/

[17] vds.cnes.fr/manifestations/PV2002/DATA/6-3_clark.pdf

[18] www.the-dma.org/cgi/dispannouncements?article=519

8.4 Will Statistical Data Remain Valuable?

Some are beginning to question whether statistical data will continue to have value, at least for traditional DSOs such as national statistical offices. The reasoning behind this concern is that vast pools of administrative data are becoming available, both from private sector sources and from various government agencies. The use of this administrative data could, at least in part, supplant the census and survey activities of statistical agencies. Will so much administrative data be readily available that statistical data will be largely irrelevant to many data user needs?

Such questions beg for a clear understanding of what is meant by statistical data. Just as with objects such as African shields, which have different meanings depending on purpose and context—say whether in a village, an ethnographic museum, or an art gallery—data need not necessarily be created for statistical purposes in order to become statistical. Administrative data can be used for statistical purposes and then it becomes by definition statistical data. As we have argued in Chapter 1, the uses of data for statistical purposes are myriad and important. Indeed in coming years they can only grow in application and significance. As we know, those entrusted with data to be used for statistical purposes have confidentiality responsibilities. Importantly, because administrative data may be required to obtain certain benefits, or even that providing the data may be legally mandated for some individuals or enterprises, a DSO intending to use such data for statistical purposes must be cognizant of the privacy issues involved in extending the usage scope beyond the original administrative purpose.

Within the United Kingdom, the Economic and Social Research Council has recently set up the Administrative Data Liaison Services, the purpose of which is both to advise users on the possibilities of accessing and reusing administrative data for statistical purposes, and to advise administrative data holding DSOs on the entailments of making such data available for statistical use. In both of these advisory activities statistical confidentiality is central.

Despite these various initiatives for reusing administrative data, we see statistical data continuing to be valuable. In the context of the Survey of Income and Program Participation (SIPP), CNSTAT (2009) recommended ways in which administrative records could be used more effectively. They raised confidentiality concerns about direct use of such records and argued for the continuation of the survey basis for SIPP. Critical in these concerns is that administration data, whilst a potentially valuable additional resource, have not been collected with such reuse in mind and therefore lacks: (1) the variable coverage that researchers want and (2) the statistical rigor in data collection processes regarding matters such as missing data and metadata processes. These issues are not insurmountable, but solving them would require a cultural change in how and why administrative data are collected. One could imagine a fusion of administrative and statistical data collection processes, perhaps extending the notion of population registers to purpose-specific administrative data. Notwithstanding such possibilities it is reasonable to say that for the foreseeable future data collected solely for statistical purposes will remain with us.

8.5 New Data Types

Traditionally, statistical data have appeared in the form of numerical measurements and assessments. Today, there is rapid growth in the collection and presentation of data in a variety of other forms and media, specifically data of the following sort:

1. *Geospatial.* Geospatial data integrate maps with numerical attributes.
2. *Audio and video.* The capacity to generate, store and search audio and video data is exploding.
3. *Biometric recognition.* The output of biometric devices, such as fingerprint scanners, is being stored in large databases.
4. *Biological material.* Tissue and blood samples, for example, are increasingly being gathered and stored for research and administrative purposes. This can provide genetic marker information.
5. *Electronic network.* With the growth of internet technology, e-traffic and other forms of communications data are being captured for research and management purposes.

To a large extent, existing confidentiality protection procedures are not adequate for these new data forms, so they present a challenge for policy makers and for confidentiality researchers. We comment in turn on confidentiality concerns for each of these five types of data.

8.5.1 Geospatial Data

Geospatial data integrate maps with numerical attributes. Newly implemented technologies such as satellite imagery, global positioning systems (GPS), and RFID tags have made locating and tracking people relatively easy and inexpensive. Importantly for social research, the linkage of spatial data with personal data has much value in developing understanding and public policy in social, economic, political, environmental and public health realms. Its use may also raise disclosure risk (National Research Council, 2007). The Socioeconomic Data and Applications Center (SEDAC) at Columbia University has prepared an online bibliography[19] on the topic of confidentiality and geospatial data. The widespread availability of GIS software allows detailed geographical specificity with both numerical and graphical components. Pickle et al. (2005) note that between 1994 and 2002 the number of publications that use GIS data for health research had grown by about 26% per year, four times the rate of increase in the number of articles on human health in general. Brownstein et al. (2006) discuss the ability to identify patients from published maps. They created a simulated map of 550 geographically coded addresses of patients in

[19] sedac.ciesin.columbia.edu/confidentiality/SelectedDocuments.html

Boston using the minimum figure resolution required for publication in the *New England Journal of Medicine*. Using standard GIS techniques they were able to precisely identify 432 of the addresses (79%) and identified all 550 addresses within 14 m of the correct address. Since this suggests a confidentiality problem, what can be done about it? A common approach has been to map according to administrative unit rather than home address. However, they note that such aggregation constrains visualization of disease patterns. They rather suggest randomly relocating patients' addresses within a given distance of their true location. More comprehensively, National Research Council (2007) discusses how to protect confidentiality for data that link social and spatial information. It suggests that some progress has been made in developing technical restricted data approaches to confidentiality protection. This progress includes measurement of disclosure risk for particular geospatial data products and quantifying data quality. It remains unclear how effective various proposed SDL methods that perturb spatial information may be in reducing disclosure risk while maintaining data utility. This leaves two alternative approaches to geospatial data dissemination—restricted access, including licensing and data enclaves, or release of synthetic data. Machanavajjhala et al. (2008) developed synthetic data to statistically mimic source data from the US Census Bureau on commuting patterns in the United States.

8.5.2 Audio and Video Data

The capacity to generate, store, and search audio and video data is exploding. Some of this is in formal research studies, such as video of psychiatric patients being used to evaluate comprehension of informed consent. Others involve use of audio and video in observational studies.

Of particular concern for confidentiality, with video capture the distinction between persons and their data is blurred. Also, it is hard to assure anonymity when there are so many visually recognizable aspects of a person that are revealed in a video. Not only facial features, but observed clothing and gait can be identifying as well. This raises new concerns for how informed consent should be obtained from subjects and what should be done if a researcher happens to recognize a subject. These are particular problems for video data archives.

8.5.3 Biometric Recognition Data

Biometric recognition data result from systems that measure certain physiological or behavioral characteristics of individuals. These systems include fingerprint scans, hand geometry, facial recognition, iris scans, retinal scans, voice recognition and signature verification. Based on such measurements, these systems seek to automatically identify, verify identity, or authenticate individuals.

Biometric recognition is being widely promoted as a useful technology in far-ranging contexts. As just one example, motivated by facts like that the nineteen

9/11 terrorists had a total of 63 valid driver's licenses, it is touted as a way to help identify terrorists. Concerned about better access control to physical facilities, the US government has put forth standards, particularly FIPS-201 (Federal Information Processing Standards Publication 201), that specify requirements for personal identity verification for Federal employees and contractors. It includes the use of biometric information that can be stored on a smart card and matched to that in an external database.

Other domains for biometric recognition include identifying crime perpetrators, patient tracking in medical informatics, and the personalization of social services. In spite of substantial efforts, there remain unresolved questions about the effectiveness and management of systems for biometric recognition, as well as their appropriateness and societal impact. From our perspective, a specific societal impact that needs to be addressed is how confidentiality can be protected in biometric recognition systems.

Biometric recognition inherently involves the use of databases both for functional reasons and administrative reasons. Functionally, biometric systems submit the results of a physiological or behavioral measurement to a search of a database containing reference or enrolled representations of the same biometric characteristics captured from a population of individuals. An example of this is national ID cards or passports that contain biometrics. Administratively, there are many evident applications such as a homeless shelter checking whether an applicant has a criminal record and law enforcement checking whether a target individual used a certain facility at a certain time. Beyond functional or administrative use of a biometric database, there are policy and research uses. These research uses generally involve sharing data by integrating records with other databases and direct dissemination to a variety of data users.

How sensitive are biometric data? In most cases they can be considered highly sensitive. If biometrics are to be useful for recognition purposes, they should be unique identifiers of an individual. Thus they have the potential for aiding in identity theft. Also, they can reveal personal characteristics acknowledged to be sensitive, for example, whether a person has a particular disease. Given these observations, how can biometric data be shared? Technical procedures need to be developed that can be used to obtain value from sharing biometric data while protecting confidentiality. Also, administrative policies must be implemented to appropriately guide biometric data sharing.

8.5.4 Biological Material Data

Medical researchers have always collected various biological specimens—blood, urine, saliva, and tissue samples—as an integral part of their work. Recently social scientists have also begun to collect certain biological specimens. In part, this has been motivated by fairly mundane issues like corroborating survey responses, such as whether a respondent is infected with a herpes virus. More significantly, this collection has been motivated by research questions that link disease or genetic makeup to behavior. For example, Huizinga et al. (2006) examine whether a monoamine oxidase A genotype may moderate the influence of adolescent maltreatment on adult antisocial behavior. Questions like this one can only be addressed when genetic and other biological data are combined with information about social and environmental factors. Biomarkers have been used to identify cardiovascular risk factors, metabolic process measures, and immune-system activity. Researchers have assessed such biological factors as Epstein–Barr virus antibodies (a marker of immune function), HDL cholesterol (indicator of cardiovascular risk), and hemoglobin A1c (a marker of glucose intolerance).

Because of the power of genetic analysis and the ability to identify individuals through their social and biological data, sharing data that link genetic information with social and behavioral data presents a high probability of disclosure. Also, the consequences of disclosure are high. According to Greely (2009), many view obtaining biomeasures to be quite different from collecting social survey information. They perceive biomeasures to be more:

- "objective"
- revealing about a person's health, even about things that may be hidden from the data subject, such as HIV status
- important to protect because of the view that health-related information should be kept private

At present, no restricted data SDL procedure appears workable, because the masking of the genetic data would have to be so extensive that the data utility would be too low. Therefore, a sensible guideline is to share only under restricted access conditions, such as licensing and data enclaves.

8.5.5 Network Data

With the advance of internet technology, e-traffic and other forms of communications data are being captured for research and management purposes. The accelerating growth of social media and online communities, such as Facebook, have generated enormous amounts of social network data that present research opportunities and confidentiality challenges at the nexus of statistics, computer science, and the various social sciences. As Kleinberg (2007) notes researchers have investigated online social systems related to communication, community formation, information-seeking and collective problem-solving, marketing, the spread of news, and the dynamics of popularity. Collecting and analyzing data on social networks has become an increasingly important branch of quantitative social science in the last 20 years (see Freeman (2006) for a review of the development). Recently some authors (e.g., Tranmer et al., 2010) have even suggested that addressing social research questions with data that do not include social network information may increase the risk of incorrect inferences being made.

However, social network data also possess significant confidentiality concerns. This problem is twofold. First, to be of any value social network data normally need to be for a population and not a sample, that is they need to be for the entire network of interest. Also perturbation of a network will tend to destroy data quality far more than say with standard microdata. So, standard techniques for protecting data do not work with social network data. Second, as Backstrom et al. (2007) have demonstrated intruding on a network by matching network fragments against the released network is relatively straightforward, particularly if the nodes (population units) in the network have associated attributes (equivalent to variables within a microdata set). Further, Kleinberg (2007) notes that "some of the richest emerging sources of social interaction data come from settings such as e-mail, instant messaging, or phone communication in which users have strong expectations of privacy." Backstrom et al. (2007) demonstrate that simply replacing personal identifiers with arbitrary codes is not sufficient to eliminate the risk of identity disclosure. Therefore, it seems likely that network data, at least that pertaining to live persons and extracted from non-public data, will need to be subject to restricted access arrangements and so stored in safe settings.

8.6 Privacy Preserving Data Mining

Over the past decade the computer science community has accelerated its research into privacy. An early review is Adam and Wortmann (1989). A significant component of this research has been under the label of privacy preserving data mining, as introduced by Agrawal and Srikant (2000). Consistent with general usage, Gehrke (2006) defines data mining as, "the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data." Data mining is privacy preserving provided the data do not violate the privacy (or confidentiality, depending on context) of the data subject. As Vaidya et al. (2006) note, the goal of privacy preserving data mining is win-win. The data user is able to do appropriate statistical analyses on the data, the DSO is protected against misuse of the data, and the privacy of the data subject is assured. The reader will note that this motivation is essentially the same as that in statistical confidentiality.

Other useful references for privacy preserving data mining include Clifton et al. (2002), Dwork et al. (2006), Dwork (2006), Evfimievski et al. (2003), Martin (2007), and Rastogi et al. (2007). Much of this research has relevance to statistical confidentiality, but has a slightly different perspective. Dwork et al. (2009) identify several unanswered questions about the implications of using privacy preserving data mining methods in the context of statistical inference.

Basic to much of the computer science research are attempts to establish privacy guarantees. This requires formal definitions of "privacy." Against these definitions various SDL methods (often called *techniques for data publishing* in the computer science literature) are tested. What is sought is a guarantee of privacy of sensitive information, even against adversaries who possess damaging background knowledge. Hopefully the transformed data remain useful for data analysis.

Machanavajjhala et al. (2008), which was discussed in Section 8.5.1 in the context of geospatial data, proposed the first formal privacy analysis of synthetic data generation—a method we examined in Section 5.13—and found "that while some existing definitions of privacy are inapplicable to our target application, others are too conservative and render the synthetic data useless since they guard against privacy breaches that are very unlikely. Moreover, the data in our target application is sparse, and none of the existing solutions are tailored to anonymize sparse data." It concludes that future research is needed on methods for incorporating outlier identification, suppression, and modeling to the privacy and utility guarantees for this mapping application.

Xiao et al. (2009) introduce CAT, the Cornell Anonymization Toolkit, for interactive SDL by identifying records that have high disclosure risk under various data snooper models. CAT allows a DSO to adjust SDL parameters and examine both disclosure risk and data utility in the anonymized data. Application is made to two-way contingency tables.

The relevance of computer science research to statistical confidentiality can only increase in the future as large databases resulting from the new data types discussed in the previous section become even more pervasive.

8.7 Other New Issues for Statistical Confidentiality

Based on recent trends we project that changes in five other areas will raise new concerns for those working to provide statistical confidentiality. These areas are technological advances, the activity of sophisticated privacy advocates, increased expectations about data access, new confidentiality legislation, and challenges in communicating confidentiality protections. We discuss each area in turn.

8.7.1 Technological Advances

Recent developments in algorithms and software for record linkage worry DSOs that SDL methods that were once adequate may no longer suffice. Indeed, some consider this concern so great that public-use microdata files may not be made available in the future. As perhaps a harbinger of this increased concern, consider this public health data situation: The National Center for Health Statistics (NCHS) made publicly available the first two releases of a linked file of the National Health Interview Survey to the National Death Index. As an example of its value, with the linked file, as Horm[20] notes, "the impact of poverty or health insurance status on the risk of dying could be explored while simultaneously controlling for age, sex, race, acute or chronic conditions." Unfortunately for broad use of this file, the third release,

[20] www.fcsm.gov/working-papers/horm.pdf

which follows both respondents from the earlier survey years and adds new survey years, is only available for restricted use in the NCHS Research Data Center. It is not publicly available.

A second area which is both a problem and opportunity for statistical data is the rise of distributed computing. The grid, cloud computing, e-science, cyberinfrastructure, and Web 2.0 are a set of interrelated ideas which are being developed in commercial and scientific contexts as we write. The precise sociotechnical meanings of these concepts will only become apparent as their development progresses. One thing that is now clear is that they will heavily influence how we access, use, manage, collect, and store data. Increasingly we will be able to seamlessly bring together, combine, and enhance data which exist in different forms and at different locations. Such capabilities create whole new challenges for statistical confidentiality.

The potential for the combination of these two areas, data linkage and distributed computing, is indicated by new software such as the Autonomy system. Here is a quote from Autonomy's website:

> Autonomy addresses the problem of information management borne out of the exponential increase in unstructured data (i.e. documents, phone conversations, presentations, video, email etc). It allows unstructured information to be personalised and organized by automatically analysing and understanding that information based on its contents.[21]

In short, the pace of technological change evident throughout the past 50 years shows no sign of abating and will continue to have significant impacts on the relationship between researchers and data. We are struck by the continued relevance of Marshall McLuhan's famous sound bite: "the medium is the message"; McLuhan (1964). For, as the reader may have noted at this point, we foresee that this technology change will not only change the way we access data but change the nature of data itself.

8.7.2 *Increased Expectations About Data Access*

More and more researchers and data analysts wield enormous computing power and sophisticated statistical methodology. These tools beg to be used on the rich, extensive data that are now collected, stored, and integrated into data warehouses. In many areas of study, data can be obtained easily through web access. No matter which DSO, researchers want rich data, and without major hassle. Given this capability of supply and interest in use, we are hardly surprised that researchers feel frustrated when access is denied or curtailed under the guise of confidentiality protection. As a consequence, researchers may subsequently fail to seek access to data that may be valuable for their research. On the other hand, rather than give up, researchers may place untoward pressure on DSOs to make a data product available—even if adequate confidentiality protection cannot be assured for this product or the product

[21] www.autonomy.com/

may, because of SDL processing, be misleading or not that useful to researchers. As Winkler (2004b) has argued in the case of public-use microdata products, DSOs need to do more to substantiate both reasonably high levels of data utility—some valid analyses can be carried out with the data—and appropriately low levels of disclosure risk—the data do not allow reidentification of particular individuals.

8.7.3 Sophisticated Privacy Advocates

Based on the experience of recent years, we project that privacy advocates will become increasingly sophisticated, particularly in their access to media outlets and their knowledge of information technology. The concerns that they express could have a negative impact on response rates in surveys. To a researcher this may seem wholly bad, but that is not the case. As Sieber (2007) notes, "... parents recruited for a study of child-rearing practices should be warned of the limits of confidentiality (for example, that there is mandatory requirement for the reporting of evidence of child abuse). While this might distort the researcher's random sampling scheme and jeopardize generalisability by eliminating those who decline to participate, it provides a higher level of confidence in the candor of those who choose to participate and suggests conducting a parallel study of parents who have been convicted of child abuse." Importantly, privacy advocates can be expected to argue for new confidentiality legislation that may further restrict researcher access to data.[22]

8.7.4 New Confidentiality Legislation

In response to privacy concerns and the reality of globalization we expect further developments to enhance privacy protections. Some of these developments will come from the private sector (as Daniel Klein has argued is best[23]) but certainly new legislation will have an impact on confidentiality, especially regarding cross-border flows of personal data. Our expectation is that such legislation will provide a more uniform set of rules governing the practices of DSOs. In the United Kingdom, the recent Statistics Registration Act was a carrot and stick affair—enshrining the principle of researcher access in legislation but at the same time introducing harsh penalties for data abuse.

8.7.5 Demand for Data from Researchers

Researchers face many difficulties when trying to find and use official data for research purposes. These difficulties range from the practical and organizational to the political, cultural and legal. The first hurdle is to find whether appropriate data

[22] www.edri.org/edri-gram/number7.2/no-voluntary-data-retention

[23] www.fee.org/publications/the-freeman/article.asp?aid=1804

are available for the regions under study, and if so, how and where the researcher can access them. Occasionally, some of these improvements have jointly taken place between national statistical offices and data archives. However, these instances are limited and, where they have occurred, the benefits tend to be limited to researchers in the country from which the data originate since there are no coherent mechanisms in place to enable cross-border data exchange. In the best-case scenarios, researchers have access to online catalogs, links to the supplying organization, and liberty to register online and acquire data delivery, but in many countries none of these facilities are available.

Researcher accreditation involves legal issues as well as practices and is considered to be one of the most significant hurdles to access. It does not even feature as an access mechanism in some countries, and even where it is in place the details differ from one country to another. This activity will work to establish a single cross-border reciprocal arrangement under which accreditation by one institution will automatically be recognized by the next. This will require high-level discussion and information exchange between all partners. Some DSOs, who own the primary responsibility for risks, are concerned that differences in accreditation procedures imply differences in standards and controls. In contrast, archives and other access providers are concerned that the procedures impose either too great a burden, or otherwise provide insufficient standards and controls. This networking activity should enable these competing views to be reconciled satisfactorily so that researchers no longer find their attempts to gain access across borders a frustrating, confusing, and burdensome experience.

8.7.6 Challenges in Communicating Confidentiality Protections

As is evident from the previous chapters, because much of the literature being developed for SDL is highly technical and requires mathematical sophistication to fully comprehend, DSOs need to employ confidentiality specialists to determine the relevance of new techniques to their needs. But to be effective, DSOs also have to work to communicate to data providers, privacy advocates, and oversight bodies such as IRBs just how these techniques provide confidentiality protection. To do this, DSOs may need to engage specialists in risk communication.[24] An effective tool in this regard can be well-structured websites (see, e.g., recommendations of Sieber (2007) for communication with IRBs, researchers, and students). The importance of this is not to be underestimated. Some DSOs tend to view wide-scale communication about statistical confidentiality as a risk in itself. Their concern is that such communication would decrease response rates. But good communication about risks and benefits could lead to greater buy-in by respondents and also reduced negative consequences if a disclosure event takes place; a public more aware that the risk is non-zero is less affected if the risk turns to reality.

[24] centerforriskcommunication.com/

8.8 Will There Be New Forms of Data Snooping?

As we noted in Chapter 1, DSOs have found it useful to frame discussion about confidentiality in terms of protection against a data snooper. Can we anticipate changes in the conceptualization of the data snooper?

8.8.1 The Data Snooper of the Future

Future data snoopers will include all the types and envisioned by Paass (1988) and Elliot and Dale (1999) discussed in Chapter 1. But new types of data snooper will likely emerge. Some will be motivated differently than previously. Some will employ currently unanticipated strategies in attempting to breach confidentiality. We anticipate that DSOs will have to learn how to cope with attacks from data snoopers who are motivated by cyberterrorism, commercial espionage, governmental intrusion, and identity theft. Let us examine these motivations in turn.

Cyberterrorism. Generally, cyberterrorism is thought of as "leveraging of a target's computers and information technology, particularly via the Internet, to cause physical, real-world harm or severe disruption."[25] Presumably to be effective as a terrorist threat, an attacker would generally have to compromise the confidentiality of a substantial number of records. It is perhaps easiest to understand how this could be achieved through a computer security breach, rather than through inferential disclosure based on a data product that a DSO may have disseminated. However, sophisticated cyberterrorists may see advantages in undermining public faith in a government's data processes. Because of cyberterrorism threats, DSOs will need to strengthen their computer security and this may also have some negative effects on data access. Staff will need to become more knowledgeable about the nature of such threats and about the technology needed to thwart them.

Governmental intrusion. Government agencies, for reasons typically justified by arguments of furthering their own mission, may seek to target personal or proprietary information held by DSOs. Whether through secretive spying or more public demands for access, such activities compromise assurances of statistical confidentiality given by a DSO. It is likely that DSOs will have to fight for legal protections of their data from such intrusions. Indeed, otherwise, legislation may move to explicitly allow such intrusions. For example, passed in the aftermath of the attacks of September 11, the USA Patriot Act of 2001 in Section 508 of Title V permitted the Department of Justice "to obtain and use for investigation and prosecution 'reports, records, and information (including individually identifiable information)' in the possession of

[25] en.wikipedia.org/wiki/Cyber-terrorism

the National Center for Education Statistics (NCES) that had hitherto been protected by the 1994 National Center for Education Statistics Act."[26]

Identity theft. The US Federal Trade Commission says, "Identity theft occurs when someone uses your personally identifying information, like your name, Social Security number, or credit card number, without your permission, to commit fraud or other crimes." It estimates that some 9 million identities are stolen each year.[27] Generally, the purpose of identity theft is to fraudulently authenticate an identity in some transaction. Usually this is thought of as strictly financial transactions, but a different and particularly noxious variant is medical identity theft. "Medical identity theft occurs when someone uses a person's name and sometimes other parts of their identity—such as insurance information—without the person's knowledge or consent to obtain medical services or goods, or uses the person's identity information to make false claims for medical services or goods. Medical identity theft frequently results in erroneous entries being put into existing medical records, and can involve the creation of fictitious medical records in the victim's name."[28] The possibility of the use of statistical disclosure to enhance the information available to the identity thief cannot be ruled out, particularly in those situations where the thief/snooper has response knowledge—that is, knows that the individual whose identity they have stolen has responded to a particular survey. The significance of this from a media perspective should be noted. The recent report of a successful reidentification study was picked up by the media as a threat in terms of fraud rather than privacy *per se*. For example, writing in the United Kingdom's *Observer* newspaper policy editor reports the study like this: "Computer security: fraud fears as scientists crack 'anonymous' datasets."[29]

Commercial espionage. Going well beyond the usual concerns about stealing intellectual property, commercial espionage could compromise confidentiality of records held by a DSO. The snooper could be seeking to gain information, for example, on key personnel of a competitor. A commercial spy might find financial or health data to be particularly valuable.

8.8.2 New Attack Modalities

Besides different motivations enlarging the cast of characters who may be motivated to be data snoopers, changing technology will likely change attack modalities. Increasingly, the way information is disseminated is through online databases.

[26] Seltzer and Anderson (2002) www.amstat.org/sections/srms/Proceedings/y2002/Files/JSM2002-000351.pdf

[27] www.ftc.gov/bcp/edu/microsites/idtheft/consumers/about-identity-theft.html

[28] www.worldprivacyforum.org/pdf/wpf?exsum?medidtheft2006.pdf

[29] www.guardian.co.uk/technology/2010/jan/24/computer-security-crime-anonymous-datasets

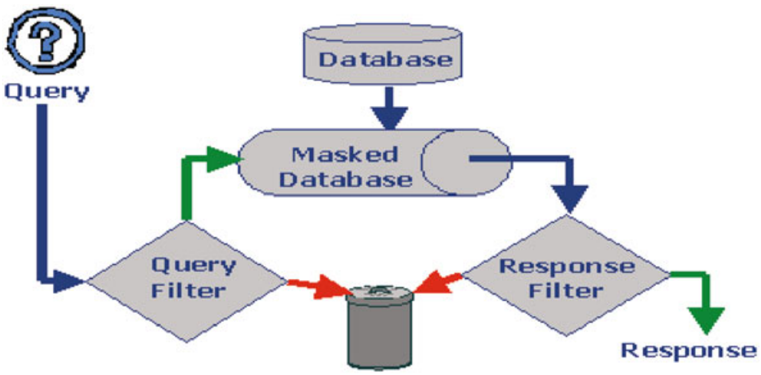


Fig. 8.1 SDL of online databases

The following are examples of websites maintained by national statistical offices: *American FactFinder*[30] (see Zayatz and Rowland, 1999), *Office of National Statistics*,[31] and *Statistics Netherlands*.[32]

With this trend for data dissemination via the web now well-established, we can expect that data snoopers will attempt to compromise confidentiality through online attacks. In the face of this threat, DSOs have several options for SDL, as Fig. 8.1 illustrates.

Referring to Fig. 8.1, query filters include, for example, restrictions on the dimensionality of a requested table, say to no more than three-dimensional tables. The database may be masked through topcoding or data swapping.

The response filter may not allow tables where the value (say, R&D expenditures) for a cell came from only one or two entities (say, privately held optical networking firms). A basic work that lays out key ideas for online disclosure limitation is Adam and Wortman (1989). Methodology for additive noise in repeated query systems is developed in Duncan and Mukherjee (2000). A variety of general concerns about remote access are laid out by Blakemore (2001). More recently there has been an upsurge in the work on the methods of automatically filtering output. See, for example, O'Keefe and Good (2008), Sparks et al. (2008), and Simard (2009).

We also can expect data snoopers to have more capability to make linkages among databases with many records and attributes. In particular, they can seek to reidentify records by linking administrative records with public-use data records (Winkler, 2004a).

[30] factfinder.census.gov/servlet/BasicFactsServlet

[31] www.statistics.gov.uk/

[32] www.cbs.nl/en/figures/keyfigures/index.htm

8.9 What New Strategies of Disclosure Limitation Should Be Developed?

Duncan and Pearson (1991) prophesized that, "Agencies will employ statistical masks that are effective yet faithful to the original data. Statistical methods for the analysis of masked data will be developed, cheaply available and easy to use." Although substantial progress has been made on developing such SDL techniques, much still remains to be done, especially with regard to online access to data.

We see the following areas as ones where some research contributions have been made, but additional work is needed to meet the pressing needs of DSOs:

1. *Obtaining a better understanding of the empirical disclosure risks that arise when data are being disseminated.* Most research to date in this area has focused on the possibilities a data snooper has to compromise confidential data. Thus, the emphasis has been on the *potential* for disclosure. Little empirical work has addressed what might motivate someone to act as a data snooper. Nor has there been much beyond speculation about what the actual harm might be to a DSO of a compromise of confidentiality. Solid empirical work in these areas will lead to better understanding of the real disclosure risk involved in the release of data products.
2. *Quantifying information loss due to the implementation of SDL procedures.* We believe that progress on this can be made using two parallel strategies. One strategy is to pursue the information-theoretic framework first put forth in Duncan and Lambert (1986). The other strategy is to empirically examine how different classes of researchers actually use data. Certain researchers, for example, those in academia with abundant computing resources and strong methodological skills, want data that are longitudinal and that have fine geographical detail. Other researchers may be content with more aggregated data. It is likely to be difficult to serve the needs of the first class of researchers with publicly available data products. Instead, they may be required to obtain the data they need under restricted access conditions.

We see three especially promising areas for future research:

1. *Virtual or synthetic data.* As we discussed in Chapter 5, synthetic data sets consist of records of individual synthetic units rather than records the agency holds for actual units. Rubin (1993) suggested synthetic data construction through a multiple imputation method. This work has been extended by Fienberg (1997), Fienberg et al. (1997, 1998), Raghunathan et al. (2003), Reiter (2003), and Polettini (2003). Work remains on how best to develop the probability model, the *synthesizer*, that is to be used through Monte Carlo methods to generate the synthetic data. Also a better understanding is needed of how to interpret synthetic data. Two situations are of particular interest: (1) the extent to which the released data set is to be synthetic (see the discussion in Little and Liu 2002,

2003 of partially synthetic data) and (2) the extent to which the synthetic data match the source data in the case of outliers. Since a (completely) synthetic data record corresponds to no particular individual, identity disclosure is impossible. However, attribute disclosure is possible when a synthetic value closely matches an extreme record value and the data snooper knows that a particular individual is an outlier on that attribute.

2. *Methods for longitudinal data.* As many have noted, for example, Benedetti et al. (2003) and the Federal Committee on Statistical Methodology (1994), adequate SDL methods for longitudinal data do not exist. This is a serious lack because of the immense value of longitudinal data in empirical policy-related research (Mackie and Bradburn, 2000). In each of these areas, the R-U confidentiality map may provide a unifying framework for consideration of tradeoffs in disclosure risk and data utility. Further, it can motivate the systematic comparison of alternative disclosure limitation methods.
3. *Online access to data.* Duncan and Pearson (1991) foretold correctly that, "Electronic gatekeepers and monitors for the remote access to, and utilization of; computer databases will be widespread." Some key issues for methodological research that are currently being addressed or should be addressed include the following:

   a. Understanding the interplay between query filters, masking of a database, and response filters and the subsequent impact on disclosure risk and data utility.
   b. Appropriate models for disclosure risk in repeated query systems.
   c. The detection of anomalous behavior on the network.
   d. R-U confidentiality maps for dynamic databases (increasingly databases will be based on real-time data captures, and so will change rapidly).
   e. Confidentiality issues for multimedia data, for example, video.
   f. Consideration of disclosure limitation procedures appropriate for a hierarchy of data users, for example, in the health care domain, some researchers may have obtained clearance to more detailed data.
   g. Problems of compatibility between overlapping disclosure limited data files.
   h. Development of data utility metrics (information loss metrics) that are tailored to particular analyses (Winkler, 2004a).
   i. Implementation of secure centers, remote access and data archive at DSOs.
   j. Improvement in the consistency of social science data archives and of the researcher's access to official statistics.

8.10 Finally, an Exciting Vision for Statistical Confidentiality

For the most part, Duncan and Pearson (1991) laid out a happy vision for the future of statistical confidentiality. But they did remark, "Although this optimistic vision is feasible, it will require substantial effort. Otherwise, a bleaker vision of the future may look like this: Agencies employ masks that make data difficult to analyze yet fail to deter data spies. Researchers are haphazardly denied access to federally collected databases, while data spies readily obtain personal information from private sources of information. Researchers, agencies and legislators spend considerable time wringing their hands about a seemingly chaotic and unprincipled and inequitable process of data access and data denial. Especially controversial data are held exclusively by federal agencies who fail to release information that may be embarrassing to government agencies, thus limiting our ability to understand these issues." Certainly these are concerns that continue today to resonate with all concerned with statistical confidentiality. In the future, surely all with responsibilities for statistical confidentiality will face the challenge of avoiding this bleak vision.

Official statistics provide a critically important backbone to many branches of empirical social science research and policy making, especially in the case of government-generated microdata and international comparative macrodata. Yet, the opportunity will continue to be there for the happier vision of using statistical confidentiality constructively to achieve the data access needed to sustain policy research and at the same time appropriately protecting those data in the interest of those who have provided them. That opportunity is exciting.

As we have continually emphasized since Chapter 1, the challenge of statistical confidentiality is to provide the tools whereby DSOs may disseminate data products of high utility to a range of bona fide users so they can address socially, economically and politically important questions without compromising the confidentiality and trust of those who provide the data. Given the many challenges identified in this chapter, even to maintain the current status of data access will require a social cultural shift involving DSOs, data users and respondents. That shift requires both fuller communication between DSOs and respondents, and a higher level of responsibility for confidentiality on the part of data users. Without these changes, respondents cannot be expected to provide accurate responses, or even any answers at all. We hope that this book will help facilitate that necessary cultural shift so that quality data can be responsibly used in addressing our society's major concerns.

Glossary

This glossary provides explanations of how particular technical terms are used in the context of statistical confidentiality. Terms generally used in broad fields of statistics, information technology or empirical research are not included. Nor do we include broad terms such as privacy and confidentiality, which are themselves major subjects of the whole of this book.

Additive table A contingency table in which the interior cells sum to the associated marginal totals. Tables can become non-additive when some perturbative disclosure limitation techniques are applied to them.

Aggregate data Data which summarize information across a population (or sample). Summary statistics, frequency tables, and magnitude tables are all examples of aggregate data.

Analysis server An Internet or LAN-based remote access system enabling the analysis of data sets to be conducted remotely.

Analytical completeness The extent to which a disclosure-limited data release allows the same analyses to be conducted as prior to the disclosure limitation being applied.

Analytical validity The extent to which a disclosure-limited data release leads to the same statistical inferences as prior to the disclosure limitation being applied.

Anonymization A process to make practically impossible the identification of subjects in a database, that is, a process of making the risk of identity disclosure negligible.

A priori knowledge Knowledge that a data snooper has which would facilitate attempts to disclose information about individual population units. This might be in the form of pre-existing knowledge about the population units, knowledge of the disclosure control process that has been applied to the data set, or response knowledge.

G.T. Duncan et al., *Statistical Confidentiality*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-7802-8,

Synonym: External knowledge

Attribute suppression Statistical disclosure limitation by eliminating all or some values of one or more attributes.

Attribution The association of information in a data set with a particular population unit.

Attribute disclosure The disclosure of information about a **population unit** without (necessarily) the identification of that population unit within a data set. This typically refers to aggregate data where something can be inferred about individuals who possess a certain combination of characteristics.

(Disclosure) Auditing The process of checking whether processed data are protected or not. When the data are tabular, then this process is done by computing the minimum and maximum value of all the risky cells. On large tables the computation of these two extreme values may be through integer linear programming.

Cell perturbation A statistical disclosure limitation method for tabular data whereby the values of some (or all) cells are replaced by some value in an interval.

Cell suppression A disclosure limitation method for tabular data whereby the values of particular cells are not released.

Confidentiality promise An assurance given to respondents that their data, once provided, will not be disseminated in a form in which they are identifiable or through which information can be learnt about them.

Confidentiality breach A situation where a confidentiality promise is broken, especially when a data snooper has actually identified a subject in the database and made use of this information.

Consent The process through which a respondent gives permission for data about them, collected by a data stewardship organization, to be used in a given way.

Controlled rounding A disclosure limitation method for tabular data whereby the values of all cells are replaced by a value in a finite set. Typically, for each cell, this set contains only two values, which are the numbers rounding up and down the original cell value to a multiple of a given base number (for example, 10). Marginal totals are maintained.

Controlled tabular adjustment (CTA) A disclosure limitation method for tabular data whereby the values of all cells are replaced by different values. It is a particular method of cell perturbation. CTA adjusts cell values to satisfy disclosure limitation ranges and tabular constraints, such as additivity, while minimizing data loss as measured by some linear measure of overall data distortion, such as the sum of the absolute values of the individual cell value adjustments. CTA replaces each risky cell by either of the two endpoints of its protection range. Certain non-risky cell values are adjusted by small amounts to restore additivity.

Cyclic perturbation A disclosure limitation method for tabular data whereby the values of all cells are replaced by different values. It is a particular technique of cell perturbation in which patterned collections of four or more cells, called data cycles, are randomly modified with some cell values increasing by one and others decreasing by one, in such a way that each row and column sum is undisturbed.

Data dissemination Any process through which (access to) a set of data are given to data users.

Data divergence The difference either between two data sets or a data set and the world in how information about a given population unit is recorded. This can be caused by coding differences, response errors, data aging, data entry errors, and other forms of misclassification.

Data protection A broad term which covers a set of legal and ethical principles and instruments which collectively delineate the ways in which data may be fairly processed by data stewardship organizations. Data protection operates at national and international levels.

Data snooper An individual, group, or organization that seeks to identify individual population units within a data set and/or discover information about a population unit, usually through a statistical linkage process of information already known to information contained within the data set.

Data Stewardship Organisation (DSO) An organization which captures, processes, holds, and/or disseminates data about population units. A DSO is under legal or ethical obligations to maintain confidentiality and provide data releases of high utility to users.

Data swapping An SDL method in which certain attribute values for some records in the source data are exchanged. For example, a sample of households may be selected and matched on a set of selected key variables with households in nearby geographic areas that have similar characteristics (such as the same number of adults and same number of children). Values of an attribute such as household income may be swapped.

Data user Any person using a data set for a bona fide, typically statistical, purpose.

Data utility The usability or research value of a given set of data. In the context of disclosure-limited data, high data utility requires both analytical completeness and analytical validity.

Data-world divergence The difference between a data set and reality. Sources of data divergence include data outdating, response errors, coding or data entry errors, differences in coding, and the effects of disclosure limitation.

Deidentification The removal of direct identifiers from a data set or individual record.

Direct identifiers Information that can identify a data subject directly, such as name or social security number. Also includes combinations of attribute values, such as address and age, that uniquely identify a data subject.

Disclosure auditing See auditing.

External knowledge Knowledge that a data snooper may have that is external to that in a particular data release.

Global recoding Any method of recoding the categories in a whole data set to a smaller set of categories.

Key variables A combination of attributes that increase disclosure risk because a data snooper may link values to those in an identified external data set.

Hierarchical data Used particularly to describe microdata which has grouping variables for two or more levels of analysis. The vast majority of microdata is geographically hierarchical. Another variant is hierarchical household data which contain all members of a set of sampled households.

Identification The association of a population unit whose identity is known with a particular microdata record.

Identification file A database, data set, or other form of data storage which contains formal identifiers for individual population units.

Identification risk The probability that a data snooper can identify a data subject in a released data product.

Identity disclosure Identity disclosure occurs when a data subject is identified from released data.

Integer linear programming Mathematical tools for optimization. Typically there is a linear objective function that must be maximized or minimized by choosing values of specified variables. These variables are subject to linear systems of equalities and/or inequalities with integer variables and possibly continuous variables.

Interval publication A disclosure limitation method for tabular data whereby the values of particular cells are replaced by intervals of values. For each value in each interval there should exist values in the other intervals such that together with the values which have not been replaced the resulting table is coherent. Each interval should contain the original value that it is replacing.

Linking The process of matching a data subject in released data to an identified subject in an external database.

Local suppression A disclosure limitation method where particular data units within a microdata file are coded as missing. This method is used within the **ARGUS** system.

Lower bound The minimum possible value of a given cell in a table of aggregate values which has been perturbed so that the exact value is not given (and possibly is even suppressed).

Macrodata Summary data, most especially data released in the form of a contingency table, hence with categorical attributes.

Magnitude tables Tabular data where each cell value has been obtained by adding the response value of all the contributors that fit within the categories of the key variables describing such cell. The response variable may be continuous or integer numbers.

Marginal total A cell value that has been computed by summing other cell values.

Matrix masking SDL methods that transform an $n \times p$ (cases by variables) data matrix Z through pre- and post-multiplication and the possible addition of noise.

Masking Any SDL method that prepares data for release by stochastic or deterministic transformation of the source data.

Mathematical programming An area of mathematics devoted to study mathematical models and algorithms to solve optimization problems. Integer linear programming is a sub-area, but there are other areas like convex optimization and quadratic programming.

Microaggregation An SDL method in which records are aggregated into groups. Instead of releasing the actual values for individual records, the mean (typically) of the group is released. Confidentiality is protected by each group having at least a minimum number of observations.

Microdata Data composed of records on individual data subjects. Each record might refer to an individual person, household, business, or other entity. The data may be directly collected for statistical purposes or obtained from other sources, such as administrative sources.

Minimal sample unique A combination of values within a data set that is unique within the data set and for which no subset of values is also unique.

Multivariate additive noise A statistical disclosure limitation method in which the released data are created by adding a random vector of disturbances to the source data. The multivariate distribution of the random vector is chosen to lower disclosure risk while maintaining adequate data utility.

Noise addition A statistical disclosure limitation method in which the released data are created by adding a random disturbance to some or all values in the source data.

***n* rule** A procedure to detect the risky cells (those requiring protection) in a table, so a procedure to solve the so-called primary problem. This procedure is a simple mechanism that takes into account only the number of contributors to each cell. Given a threshold value n, a cell is classified as “risky” when the number of contributors to this cell is less than or equal to n.

Parallel divergence A situation where two data sets both contain the same value for a given population unit/variable but both differ from a veridical representation of that population unit/variable. For example, somebody who lies to the tax collector about their income may well also lie to the census agency.

Perturbation Any disclosure limitation process where values for given data units are changed either systematically or randomly.

Primary problem Checking whether there is something to protect or not in a dataset. When there is something that needs protection, it must be individuated. For example, given a table, the primary problem is the problem of finding the cells (if any exist) that will need protection. The cells individuated by solving the Primary Problem are called "Primary Cells" (or risky cells). The secondary problem is protecting the risky cells.

Poisson model A model based on the assumption that the counts in a contingency table are Poisson distributed.

Population uniqueness The proportion of population units, within a given population, which has a unique combination of values, for a given set of variables.

Population unit A socioeconomic entity about which data may be collected and which is a member of a set of such entities. The set is referred to as the population. Entities may be individuals, households, families, businesses, or other organizations.

***p*/*q* rule** It is a procedure to detect which cells require protection in a table. In other words, it is a procedure to solve the so-called primary problem. This procedure is a mechanism more elaborated than the so-called *N* rule. In particular, the *p*/*q* rule takes into account the number of contributors to each cell and also the value of the major contributors.

Post-tabular SDL Statistical disclosure limitation processes applied to dataset once it has been aggregated into tabular format.

Pre-tabular SDL Statistical disclosure limitation processes applied to dataset prior to it being aggregated into tabular format.

Primary suppressions The cells individuated by a rule to solve the Primary Problem on a table. They are called suppressions when the method to solve the Secondary Problem is the cell suppression technique.

Probabilistic record linkage Record linkage processes where non-exact linkages are allowed.

Random rounding A statistical disclosure limitation method for tabular data that rounds each cell of a table up or down to a multiple of a given base number (say 5). It makes use of a probability function to decide the direction of the rounding for each value. For example, a cell value i will go down with probability $(1-i/5)$ and up with probability $i/5$. The main advantages of random rounding are the simplicity of the approach and the unbiased feature of the resulting table. The main disadvantage is that the resulting table may not be additive when the marginal cells are also modified with the same approach.

Random unique A record which is not special unique on a given set of variables.

Record linkage The process of linking records on different data sets. This can be done by researchers for bona fide analytical reasons or by a data snooper attempting to link identification information to an anonymized data set.

Record suppression A disclosure control process whereby whole records are removed from a data set before dissemination (or, in the cases of samples from census data, deliberately not sampled).

Reidentification The process of determining the identity of data subjects in data releases that have been subject to statistical disclosure limitation, that is, have been deidentified.

Research Data Center A physical facilility maintained by a data stewardship organization where data users may make use of the data for statistical purposes under restricted access conditions.

Response knowledge Knowledge that a particular population unit is represented in a given data set. This usually is in the form of knowledge that the population unit responded to the survey. Response knowledge in respect of a sample survey data set can greatly increase the risk of disclosure.

Restricted access An approach to disclosure limitation in which data access is controlled by the data stewardship organization. Access controls can be physical, technical or administrative.

Restricted data An approach to disclosure limitation in which the data are transformed, reduced, or otherwise masked and could be used on its own or in combination with restricted access.

Risk metrics Numerical indices (possibly estimated probabilities of identification or attribution) designed to indicate the likelihood of disclosure events under a given set of assumptions.

Risky cells Cells in a table that have been deemed to have high disclosure risk.

Rounding A method of disclosure limitation whereby values are rounded to a particular base. This is usually applied to frequencies within a table of counts.

R-U confidentiality map A graph of the trajectory in (disclosure risk, data utility) space as the extent of statistical disclosure protection of a given method increases.

Safe setting A form of restricted access where the data analysis environment is highly controlled. This will typically be within the offices of the data stewardship organization or in satellite offices, such as research data centers, where it has assurance that confidentiality standards are maintained.

Sample uniqueness The proportion of records, within a sample data set, which have a unique combination of values, for a given set of variables.

Sensitive cells See risky cells

Secondary problem The problem that must be solved after the Primary Problem has been solved and has generated a set of risky cells. The Secondary Problem consists of protecting the risky cells. The way to proceed depends on the chosen method (like cell suppression and controlled rounding).

Secondary suppressions Used in disclosure limitation for aggregate data these are cell suppressions that are necessary in order to stop a data snooper (or data user) from recovering the original cell values. They are the output of the Secondary Problem when the chosen method is cell suppression. *"Complementary suppressions" is a synonym.*

Sensitivity A measure, using qualitative and often subjective, of the damage that would be caused (usually to the respondent concerned) if a given piece of information is disclosed about either a given population unit or an imaginary "typical" population unit.

Social acceptability A conceptual component of sensitivity which considers how responses to surveys might vary in their compliance with social norms and expectations. Consider, for example, a survey of sexual practices or attitudes toward political extremism.

Special unique A record within a microdata set which is sample unique on a given set of variables and also on a subset of those variables.

Special unique detection algorithm A method for calculating and grading the disclosure risk for individual-level categorical microdata by determining the number and scale of special unique combinations of variables for each record within a data set within some defined key.

Statistical disclosure The process by which information is inferred about individual population units.

Statistical disclosure limitation (SDL) The process of lowering the disclosure risk of data releases intended for statistical analysis purposes. *"Statistical disclosure control" is a synonym.*

Subtraction A method of attacking a table-aggregated dataset (typically a table of counts) by removing population units for which data are already known for the dimensions of the table and thereby reducing the table to a residual, possibly disclosive, form.

Swapping key A set of key variables used for pairing records in record swapping disclosure limitation regimes.

Swapping partner Records within a data swapping system which are paired on the basis of having the same (or possibly similar) values on a swapping key.

Synthetic data Data to be released which have been probabilistically generated based on a model estimated from the source data.

Target variables Variables within a data set deemed likely to be of interest to an intruder. Target variables will generally be (i) unavailable to an intruder and (ii) sensitive in some way.

Topcoding An SDL method used with interval-scale or many-category ordinal variables where values above a certain threshold are revealed only in aggregate form. Age and income often are topcoded, with only an indication that the values are above the threshold.

Unrestricted access A data dissemination process whereby no limits are placed on access nor is such access monitored. This now is mainly used for the delivery of aggregate statistics over the web.

Upper bound The maximum possible value of a given cell in a table of aggregate values which has been perturbed so that the exact value is not given (and possibly is suppressed).

References

Abowd, J.M., Woodcock, S.D.: Disclosure limitation in longitudinal linked data. In: Doyle, P. et al. (eds.) Confidentiality, Disclosure, and Data Access, pp. 135–166. North Holland, Amsterdam (2002)

Abowd, J.M., Woodcock, S.D.: Multiply-imputing confidential characteristics and file links in longitudinal linked data. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases 2004, pp. 290–297. Springer, New York, NY (2004)

Acquisti, A.: Security of personal information and privacy: technological solutions and economic incentives. In: Camp, J., Lewis, R. (eds.) The Economics of Information Security, pp. 179–186. Kluwer, Dordrecht (2004)

Adam, N.R., Wortmann, J.C.: Security-control methods for statistical databases: a comparative study. ACM Comput. Surv. **21**, 515–556 (1989)

Agrawal, R., Srikant, R.: Privacy-preserving data mining. Proceedings of the 2000 ACM SIGMOD on Management of Data, Dallas, TX, 15–18 May 2000

Bacharach, M.: Matrix rounding problem. Manage. Sci. **9**, 732–742 (1966)

Bacher, J., Brand, R., Bender, S.: Re-identifying register data by survey data using cluster analysis: an empirical study. Int. J. Uncertainty Fuzziness Knowl.-Based Syst. **10**(5), 589–608 (2002)

Backstrom, L., Dwork, C., Kleinberg, J.: Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. Proceedings of the 16th International Conference on World Wide Web, Banff, AB, 08–12 May 2007

Barabba, V.P., Kaplan, D.L.: U.S. Census Bureau Statistical Techniques to Prevent Disclosure – The Right to Privacy vs. the Need to Know. Paper presented at the 40th Session of the international Statistical Institute, Warsaw, 1975

Bayardo, R.J., Agrawal, R.: Data privacy through optimal K-anonymization. Proceedings of the 21st International Conference on Data Engineering, 2005. ICDE 2005. Tokyo (2005)

Béland, Y.: Release of public use microdata files for NPHS? mission. . .partially Accomplished! In: Proceedings of the Survey Research Methods Section, pp. 404–409. American Statistical Association, Baltimore (1999)

Benedetti, R., Franconi, L., Capobianchi, A.: Individual risk of disclosure using sampling design information. Istat Contributi n. 14/2003. Available at http://www.istat.it/dati/pubbsci/contributi/Contr_anno2003.htm (2003)

Bernardinelli, L., Montomoli, C.: Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. Stat. Med. **11**, 983–1007 (1992)

Besag, J.: Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B **36**, 192–236 (1974)

Besag, J., York, J., Mollié, A.: Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Stat. Math. **43**, 1–59 (1991)

Bethlehem, J.G., Keller, W.J., Pannekoek, J.: Disclosure control of microdata. J. Am. Stat. Assoc. **85**, 38–45 (1990)

Blakemore, M.: The potential and perils of remote access. In: Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (eds.) Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies, pp. 315–337. Elsevier, Amsterdam (2001)

Blien, U., Wirth, H., Müller, M.: Disclosure risk for microdata stemming from official statistics. Stat. Neerl. **46**, 69–82 (1992)

Boruch, R.F., Cecil, J.S.: Assuring the Confidentiality of Social Research Data, University of Pennsylvania Press, Philadelphia, PA 1979

Brand, R.: Microdata protection through noise addition. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. Lecture Notes in Computer Science, **vol. 2316**, pp. 97–116. Springer, Berlin, Heidelberg (2002a)

Brand, R.: Tests of the applicability of Sullivan's algorithm to synthetic data and real business data in official statistics. European Project IST-2000-25069 CASC, Deliverable 1.1-D1, http://neon.vb.cbs.nl/casc (2002b)

Brin, D.: The Transparent Society. Addison-Wesley, Reading, MA (1998)

Brownstein, J.S., Cassa, C.A., Mandl, K.D.: No place to hide – reverse identification of patients from published maps. New Engl. J. Med. 19 October **355**(16), 1741–1742 (2006)

Butz, W.P.: Data confidentiality and public perceptions: the case of the European censuses. American Statistical Association 1985 Proceedings of the Section on Survey Research Methods. American Statistical Association, Washington, DC (1985)

Butz, W.P., Scarr, H.A.: The 1987 German census: a trip report. Report prepared for the U.S. Bureau of the Census, Washington, DC, June 1987

Carter, R., Boudreau, J.-R., Briggs, M.: Analysis of the risk of disclosure for census microdata. Statistics Canada Working Paper, Statistics Canada, Ottawa, ON 1991

Carvalho, F.D., Dellaert, N.P., Osório, M.S.: Statistical disclosure in two-dimensional tables: general tables. J. Am. Stat. Assoc. **89**, 1547–1557 (1994)

Causey, B.D., Cox, L.H., Ernst, L.R.: Applications of transportation theory to statistical problems. J. Am. Stat. Assoc. **80**, 903–909 (1985)

Cecil, J.S.: Confidentiality legislation and the United States federal statistical system. J. Off. Stat. **9**(2), 5 (1993)

Chowdhury, S.D., Duncan, G.T., Krishnan, R., Roehrig, S.F., Mukherjee, S.: Disclosure detection in multivariate categorical databases: auditing confidentiality protection through two new matrix operators. Manage. Sci. **45**, 1710–1723 (1999)

Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., Zhu, M.Y.: Tools for privacy preserving data mining. SIGKDD Explor. **4**(2), 28–34 (2002)

CNSTAT: Reengineering the survey of income and program participation committee on national statistics. Citro, C.F., Scholz, J.K. (eds.). National Academies Press, Washington, DC (2009)

Conway, R., Strip, D.: Selective partial access to a database. Proceedings of the ACM Annual Conference, New York, NY 1976

Cox, L.H.: Suppression methodology and statistical disclosure control. J. Am. Stat. Assoc. **75**, 377–385 (1980)

Cox, L.H.: Linear sensitivity measures and statistical disclosure control. J. Stat. Plann. Inference **5**, 153–164 (1981)

Cox, L.H.: Network models for complementary cell suppression. J. Am. Stat. Assoc. **90**, 1453–1462 (1995)

Cox, L.H., Kelly, J.P., Patil, R.: Balancing quality and confidentiality for multivariate tabular data. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science, **vol. 3050**, pp. 87–98. Springer, New York, NY (2004)

Dale, A.: Confidentiality of official statistics: an excuse for privacy. In: Dorling, D., Simpson, S. (eds.) Statistics in Society, pp. 29–37. Arnold, London (1998)

Dale, A., Elliot, M.: Proposals for 2001 SARS: an assessment of disclosure risk. J. R. Stat. Soc. Ser. A **164**(part 3), 427–447 (2001)

Dale, A., Marsh, C.: The 1991 Census Users' Guide. HMSO, London (1993)

Dalenius, T.: Finding a needle in a haystack – or identifying anonymous census records. J. Official Stat. **2**(3), 329–336 (1986)

Dalenius, T.: Controlling invasion of privacy in surveys. Department of Development and Research, Statistics, Sweden (1988)

Dalenius, T., Reiss, S.P.: Data-swapping: a technique for disclosure control (extended abstract). American Statistical Association, Proceedings of the Section on Survey Research Methods, Washington, DC, pp. 191–194 1978

Dalenius, T., Reiss, S.P.: Data-swapping: a technique for disclosure control. J. Stat. Plann. Inference **6**, 73–85 (1982)

Dandekar, R., Domingo-Ferrer, J., Sebé, F.: LHS-based hybrid microdata vs. rank swapping and microaggregation for numeric microdata protection. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. Lecture Notes in Computer Science, **vol. 2316**, pp. 153–162. Springer, Berlin, Heidelberg (2002)

De Waal, T., Willenborg, L.C.R.J.: A view on statistical disclosure for microdata. Surv. Methodol. **22**(1), 95–103 (1996)

Defays, D., Anwar, M.N.: Masking microdata using micro-aggregation. J. Official Stat. **14**, 449–461 (1998)

Defays, D., Nanopoulos, P.: Panels of enterprises and confidentiality: the small aggregates method. Proceedings of 1992 Symposium on Design and Analysis of Longitudinal Surveys, pp. 195–204. Statistics Canada, Ottawa, ON (1993)

Dellaert, N.P., Luijten, W.A.: Statistical disclosure in general three-dimensional tables. Stat. Neerl. **53**, 197–221 (1999)

DeWaal, A.G., Willenborg, L.C.R.J.: Global recodings and local suppressions in microdata sets. Proceedings of Statistics Canada Symposium’ 95, pp. 121–132. Statistics Canada, Ottawa, ON (1995)

DeWaal, A.G., Willenborg, L.C.R.J.: Information loss through global recoding and local suppression. Netherlands Official Stat. **14**, 17–20 (1999). Special issue on SDC

Domingo-Ferrer, J.: On the complexity of micro aggregation. Paper presented at the UNECE Workshop on Statistical Data Editing, Skopje, Macedonia, May 2001

Domingo-Ferrer, J. (ed.): Inference Control in Statistical Databases. Springer, New York, NY (2002)

Domingo-Ferrer, J., Mateo-Sanz, J.M.: An empirical comparison of SDC methods for continuous microdata in terms of information loss and re-identification risk. Paper Presented at the UNECE Workshop on Statistical Data Editing, Skopje, Macedonia, May 2001

Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans. Knowl. Data Eng. **14**(1), 189–201 (2002)

Domingo-Ferrer, J., Mateo-Sanz, J.M., Torra. V.: Comparing SDC methods for microdata on the basis of information loss and disclosure risk. Pre-proceedings of ETK-NTTS’2001, vol. 2, pp. 807–826. Eurostat, Luxemburg (2001)

Domingo-Ferrer, J., Mateo-Sanz, J.M., Oganian, A., Torres, A.: On the security of microaggregation with individual ranking: analytical attacks. Int. J. Uncertainty Fuzziness Knowl.-Based Syst. **10**(5), 477–492 (2002)

Domingo-Ferrer, J., Sebé, F., Castella, J.: On the security of noise addition for privacy in statistical databases. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science, **vol. 3050**, pp. 149–161. Springer, Berlin, Heidelberg (2004)

Domingo-Ferrer, J., Torra, V.: A quantitative comparison of disclosure control methods for microdata. In: Zayatz, L., Doyle, P., Theeuwes, J., Lane, J. (eds.) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 111–133. North-Holland, Amsterdam (2001)

Domingo-Ferrer, J., Torra, V.: Validating distance-based record linkage with probabilistic record linkage. Lecture Notes in Computer Science **2504**, 207–215 (2002)

Domingo-Ferrer, J., Torra, V.: Disclosure risk assessment in statistical microdata protection via advanced record linkage. Stat. Comput. **13**, 343–354 (2003)

Domingo-Ferrer, J., Torra, V.: Disclosure risk assessment in statistical data protection. J. Comput. Appl. Math. **164–165**, 285–293 (2004)

Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining Knowl. Discov. **11**(2), 195–212 (2005)

Domingo-Ferrer, J., Torra, V. Mateo-Sanz, J.M., Sebé. F.: Research data center-based confidentiality research: systematic measures of re-identification risk based on the probabilistic links of the partially synthetic data back to the original microdata. Final Report technical report, Rovira i Virgili University and IIIACSIC, Catalonia, Spain 2005

Domingo-Ferrer, J., Torra, V. Mateo-Sanz, J.M., Sebé, F.: Empirical disclosure risk assessment of the ipso synthetic data generators. In: Santos, M.J., Bujnowska, A. (eds.) Monographs in Official Statistics-Work Session on Statistical Data Confidentiality, pp. 227–238. Eurostat, Luxemburg (2006)

Doyle, P., Lane, J., Theeuwes, J., Zayatz, L.: Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. Elsevier Science, Amsterdam (2001)

Duncan, G.T.: Confidentiality and statistical disclosure limitation. In: Smelser, N.J., Baltes, P.B. (eds.) International Encyclopedia of the Social and Behavioral Sciences, pp. 2521–2525. Elsevier Ltd., Oxford (2001)

Duncan, G.T.: Appendix E: Confidentiality and data access issues for institutional review boards. In: Citro, C.F., Ilgen, D.R., Marrett, C.B. (eds.) Protecting Participants and Facilitating Social and Behavioral Research, pp. 235–252. National Research Council, Washington, DC (2003)

Duncan, G.T.: Exploring the tension between privacy and the social benefits of governmental databases. In: Podesta, J., Shane, P.M., Leone, R.C.A (eds.) Little Knowledge: Privacy, Security, and Public Information after September 11, pp. 71–88. The Century Foundation, New York, NY (2004)

Duncan, G.T.: Privacy by design. Science **317**, 1178–1179. August 31 (2007)

Duncan, G.T., Chowdury, S.D., Krishnan, R., Roehrig, S.F., Mukerjee, S.: Disclosure detection in multivariate categorical databases: auditing confidentiality protection through two new matrix operators. Manage. Sci. **45**, 1710–1723 (1999)

Duncan, G.T., Fienberg, S.E.: Obtaining information while preserving privacy: a Markov perturbation method for tabular data. Eurostat. Proceedings of Statistical Data Protection ′98, Lisbon, pp. 351–362 (1999)

Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R., Roehrig, S.F.: Disclosure limitation methods and information loss for tabular data. In: Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (eds.) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 135–166. North-Holland, Amsterdam (2001)

Duncan, G.T., Jabine, T.B., de Wolf, V.A. (eds.): Panel on Confidentiality and Data Access, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council and the Social Science Research Council, Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. National Academy of Sciences, Washington, DC (1993)

Duncan, G.T., Keller-McNulty, S.A., Stokes, S.L.: Disclosure risk vs. data utility: the R-U confidentiality map. Technical report LA-UR-01-6428, Los Alamos National Laboratory, Los Alamos, NM 2001

Duncan, G.T., Lambert, D.: Disclosure-limited data dissemination (with discussion). J. Am. Stat. Assoc. **81**(393), 10–28 (1986)

Duncan, G.T., Lambert, D.: The risk of disclosure for microdata. J. Bus. Econ. Stat. **7**, 207–217 (1989)

Duncan, G.T., Mukherjee, S.: Microdata disclosure limitation in statistical databases: query size and random sample query control. Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy, Oakland, CA, pp. 278–287 (1991)

Duncan, G.T., Mukherjee, S.: Optimal disclosure limitation strategy in statistical databases: deterring tracker attacks through additive noise. J. Am. Stat. Assoc. **95**, 720–729 (2000)

Duncan, G.T., Pearson, R.W.: Enhancing access to microdata while protecting confidentiality: prospects for the future (with discussion). Stat. Sci. **6**, 219–239 (1991)

Duncan, G., Pearson, R., Jabine, T.: The pursuit of knowledge and the protection of privacy: conflicts between access to and confidentiality of surveys of U.S. doctorates. Items (Social Science Research Council), pp. 65–70 September 1989

Duncan, G.T., Roehrig, S.F.: Reconciling information privacy and information access in a globalized technology society. In: Erickson, J. (ed.) Database Technologies: Concepts, Methodologies, Tools, and Applications, pp. 1823–1843. IGI Global, Hershey, PA (2007)

Duncan, G.T., Stokes, S.L.: Disclosure risk vs. data utility: the R-U confidentiality map as applied to topcoding. Chance **17**(3), 16–20 (2004)

Dwork, C.: Ask a better question, get a better answer a new approach to private data analysis. In: Schwentik, T., Suziu, D. (eds.) 11th Proceedings of the International Conference on Database Theory, ICDT 2007, January 10–12, 2007. Lecture Notes in Computer Science, **vol. 4353**, pp. 18–27. Barcelona, Spain (2006)

Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. Proceedings of the 3rd International Conference on Very Large Data Bases, Tokyo, Japan, pp. 265–284 (2006)

Dwork, C., McSherry, F., Talwar, K.: Differentially Private Marginals Release with Mutual Consistency and Error Independent of Sample Size. Proceedings of UNECE worksession on statistical confidentiality, December 2007, pp 193–204. Manchester (2009)

Elamir, E., Skinner, C.J.: Record level measures of disclosure risk for survey microdata. J. Official Stat. **22**, 525–539 (2006)

Elliot, M.J.: DIS: a new approach to the measurement of statistical disclosure risk. Risk Manage. Int. J. **2**(4), 39–48 (2000)

Elliot, M.J.: Data intrusion simulation: advances and a vision for the future of disclosure control. Stat. J. United Nations **17**, 1–9 (2001)

Elliot, M.J.: Statistical disclosure control. In: Kempf-Leonard, K. (ed.) Encyclopedia of Social Measurement, vol. 3, pp. 663–670. Elsevier, New York, NY (2005)

Elliot, M.J.: Data Citizenship: a 21st century solution to a 20th Century problem. Keynote speech to Exploiting Existing Data for Health Research, St Andrews September (2007)

Elliot, M.J.: Privacy, Identity and Disclosure. Keynote speech to the International Conference on Communication, Computing and Security. Rourkela, February 2011. http://www.nitrkl.ac.in/conference/conference_welcome.asp?cid=30 (2011)

Elliot, M.J., Dale, A.: Disclosure risk for microdata. End of Framework IV project report to the European Union (1998)

Elliot, M.J., Dale, A.: Scenarios of attack: a data intruder's perspective on statistical disclosure risk. Netherlands Official Stat. **14**, 6–10 (1999)

Elliot, M.J., Manning, A.M., Ford, R.W.: A computational algorithm for handling the special uniques problem. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **5**(10), 493–509 (2002)

Elliot, M.J., Manning, A., Mayes, K., Gurd, J., Bane, M.: SUDA: a program for detecting special uniques. Proceeding of UNECE Work Session on Statistical Data Confidentiality, Geneva, 9–11 November 2005

Elliot, M.J., Skinner, C.J., Dale, A.: Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk. Res. Official Stat. **1**(2), 53–68 (1998)

Eurostat Manual of Business Statistics: unstats.un.org/unsd/EconStatKB/Attachment252.aspx (2005)

Eurostat: 'Manual on Disclosure Control Methods'; Produced by B. Helmpecht and D. Schackis. Office for Official Publications of the European Communities, Luxembourg (1996)

Evans, T., Zayatz, L., Slanta, J.: Using noise for disclosure limitation of establishment tabular data. J. Official Stat. **14**(4), 537–551 (1998)

Evfimievski, A., Gehrke, J., Srikant, R.: Limiting privacy breaches in privacy preserving data mining. Proceedings of the 22nd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 2006), San Diego, CA, June 2003

Federal Committee on Statistical Methodology: Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology, U.S. Office of Management and Budget, Washington, DC 1994

Federal Committee on Statistical Methodology: Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology, U.S. Office of Management and Budget, Washington, DC 2005

Fellegi, I.P.: On the question of statistical confidentiality. J. Am. Stat. Assoc. **67**, 7–18 (1972)

Fellegi, I.P., Sunter, A.B.: A theory for record linkage. J. Am. Stat. Assoc. **64**, 1183–1210 (1969)

Fienberg, S.E.: Conflicts between the needs for access to statistical information and demands for confidentiality. Technical report #577, Department of Statistics, Carnegie Mellon University. Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute, Dublin 1993

Fienberg, S.E.: Conflicts between the needs for access to statistical information and demands for confidentiality. J. Official Stat. **10**, 115–132 (1994)

Fienberg, S.E.: Confidentiality and disclosure limitation methodology: challenges for national statistics and statistical research. Commissioned by Committee on National Statistics of the National Academy of Sciences (1997)

Fienberg, S.E.: Confidentiality and disclosure limitation. In: Kempf-Leonard, K. (ed.) Encyclopedia of Social Measurement, pp. 463–469. Elsevier, New York, NY (2005)

Fienberg, S.E., McIntyre, J.: Data swapping: variations on a theme by Dalenius and Reiss. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. Lecture Notes in Computer Science, **vol. 3050**, pp. 14–29. Springer, Berlin, Heidelberg (2004)

Fienberg, S.E., Makov, U.E., Sanil, A.P.: A Bayesian approach to data disclosure: optimal intruder behavior for continuous data. J. Official Stat. **14**, 75–89 (1997)

Fienberg, S.E., Makov, U.E., Steel, R.J.: Disclosure limitation using perturbation and related methods for categorical data. J. Official Stat. **14**, 485–502 (1998)

Fienberg, S.E., Steele, R.J., Makov, U.E.: Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: data swapping and loglinear models. Proceedings of Bureau of the Census 1996 Annual Research Conference, US Bureau of the Census, Washington, DC, pp. 87–105 1996

Fisch, K., McLeod, S.: Did you know 2.0 US Chamber of Commerce, Washington, DC, June (2007)

Fischetti, M., Salazar, J.J.: Computational experience with the controlled rounding problem in statistical disclosure control. J. Official Stat. **14**(4), 553–565 (1998)

Fischetti, M., Salazar, J.J.: Models and algorithms for the 2-dimensional cell suppression problem in statistical disclosure control. Math. Program. **84**, 283–312 (1999)

Fischetti, M., Salazar, J.J.: Models and algorithms for optimizing cell suppression in tabular data with linear constraints. J. Am. Stat. Assoc. **95**(451), 916–928 (2000)

Fischetti, M., Salazar, J.J.: Partial cell suppression: a new methodology for statistical disclosure control. Stat. Comput. **13**, 13–21 (2003)

Foster, L., Haltiwanger, J., Krizan, C.J.: Aggregate productivity growth: lessons from microeconomic evidence. In: Hulten, C.R., Dean, E.R., Harper, M.J. (eds.) New Developments in Productivity Analysis, pp. 303–363. University of Chicago Press, Chicago, IL (2001)

Franconi, L., Stander, J.: Model based disclosure limitation for business microdata. The Statistician **51**, 51–61 (2002)

Franconi, L., Stander, J.: Spatial and non-spatial model-based protection for the release of business microdata. Stat. Comput. **13**, 295–305 (2003)

Freeman, L.: The Development of Social Network Analysis. Empirical Press, Vancouver, BC (2006)

Fuller, W.A.: Masking procedures for microdata disclosure limitation. J. Official Stat. **9**, 383–406 (1993)

Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman, New York, NY. ISBN 0-7167-1045-5 (1979)

Garfinkel, R., Gopal, R., Goes, P.: Privacy protection of binary confidential data against deterministic, stochastic, and insider attack. Manage. Sci. **48**, 749–764 (2002)

Gehrke, J.: Models and methods for privacy-preserving data publishing and analysis (tutorial slides). Twelfth Annual SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2006), Philadelphia 2006

Geurts, J.: Heuristics for cell suppression in tables. Master's thesis, Netherlands Central Bureau of Statistics, Voorburg (1992)

Gill, L.: Methods for automatic record matching and linking and their use in national statistics. National Statistics Methodology Series, no. 25. Office for National Statistics, London (2001)

Goldstein, H.: Multilevel Models in Educational and Social Research. Charles Griffin, London (1987)

Gomatam, S., Larsen, M.D.: Record linkage and counterterrorism. Chance **17**(1), 25–29 (2004)

Gomatam, S., Karr, A.F., Sanil, A.P.: Data Swapping as a Decision Problem. J. Off. Stat. **21**(4), 635–655 (2005)

Greely H., Collecting biomeasures in the panel study of income dynamics: ethical and legal concerns. Biodemography Soc. Biol. **55**(2), 270–288 (2009)

Greenberg, B.V.: Disclosure avoidance research at the census Bureau. Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, DC, pp. 144–166 1990

Greenburg, B.V., Voshell, L.: The geographic component of disclosure risk for microdata. SRD Research report Census/SRD/RR-90/13, Bureau of the Census, Washington, DC 1990

Griffin, R.A., Navarro, A., Flores-Baez, L.: Disclosure avoidance for the 1990 Census. Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 516–521 (1989)

Hansen, S.L., Mukherjee, S.: A polynomial algorithm for optimal univariate microaggregation. IEEE Trans. Knowl. Data Eng. **15**(4), 1043–1044 (2003)

Heitzig, J.: The Jackknife method: confidentiality protection for complex statistical analyses. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Geneva, Switzerland, 9–11 November (2005)

Huizinga, D., et al.: Childhood maltreatment, subsequent antisocial behavior, and the role of monoamine oxidase A genotype. Biol. Psychiatry **60**(7), 677–683 1 Oct (2006)

Jabine, T.B.: Procedures for restricted data access. J. Official Stat. **9**(2), 537–589 (1993a)

Jabine, T.B.: Statistical disclosure limitation practices of United States statistical agencies. J. Official Stat. **9**(2), 427–454 (1993b)

Jaro, M.A.: Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J. Am. Stat. Assoc. **84**(406), 414–420 (1989)

Jewett, R.: Disclosure analysis for the 1992 economic census. Technical report, U.S. Bureau of the Census, Washington, DC 1993

Kamlet, M.S., Klepper, S., Frank, R.G.: Mixing micro and macro data: statistical issues and implication for data collection and reporting. Proceedings of the 1983 Public Health Conference on Records and Statistics, U.S. Department of Health and Human Services, Hyattsville, MD 1985

Keller, W.J., Bethlehem, J.G.: Disclosure protection of microdata: problems and solutions. Stat. Neerl. **46**, 5–19 (1992)

Kelly, J.P.: Confidentiality protection in two and three-dimensional tables. Ph. D. thesis, University of Maryland, College Park, MD (1990)

Kelly, J.P., Golden, B.L., Assad, A.A.: Using simulated annealing to solve controlled rounding problems. ORSA J. Comput. **2**, 174–185 (1990)

Kelly, J.P., Golden, B.L., Assad, A.A.: Cell suppression: disclosure protection for sensitive tabular data. Networks **22**, 397–417 (1992)

Kelly, J.P., Golden, B.L., Assad, A.A.: Large-scale controlled rounding using tabu search with strategic oscillation. Ann. Oper. Res. **41**, 69–84 (1993)

Kennickell, A.B.: Multiple imputation and disclosure control: the case of the 1995 survey of consumer finances. In: Alvey, W., Jamerson, B., (eds.) Record Linkage Techniques 1997, pp. 248–267. National Academy Press, Washington, DC. http://www.fcsm.gov (1999)

Kennickell, A.B., Lane, J.: Measuring the impact of data protection techniques on data utility: evidence from the survey of consumer finances. In: Domingo-Ferrer, J. (ed.) Privacy in Statistical Databases. Lecture Notes in Computer Science, **vol. 4302**, pp. 291–303. Springer, New York, NY (2006)

Kleinberg, J.M.: Challenges in mining social network data: processes, privacy, and paradoxes. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, CA, 12–15 August, pp. 4–5 2007

Kim, J.J.: A method for limiting disclosure in microdata based on random noise and transformation. Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 370–374 1986

Kim, J.J., Winkler, W.E.: Masking microdata files. Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 114–119 (1995)

Kim, J.J., Winkler, W.E.: Multiplicative noise for masking continuous data. Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA 2001

Kirkwood, C.W.: Strategic Decision Making: Multi objective Decision Analysis with Spreadsheets. Duxbury Press, Belmont, CA, (1996)

Kooiman, P., Nobel, J., Willenborg, L.: Statistical data protection at statistics Netherlands. Netherlands Official Stat. **14**, 21–25 (1999)

Kirkwood, C.W.: Strategic Decision Making: Multiobjective Decision Analysis with Spreadsheets. Duxbury Press, Belmont, CA (1996)

Lambert, D.: Measures of disclosure risk and harm. J. Official Stat. **9**, 313–331 (1993)

Lane, J.: Key Issues in Confidentiality Research: Results of an NSF Workshop. http://www.nsf.gov/sbe/ses/mms/nsfworkshop_summary1.pdf (2003)

Laszlo, M., Mukherjee, S.: Minimum spanning tree partitioning algorithm for microaggregation. IEEE Trans. Knowl. Data Eng. **17**(7), 902–911 (2005)

Lee, J., McClellan, M.B., Skinner, J.S.: The distributional effects of medicare (January 1999). NBER Working Paper No. W6910 (1999)

Little, R.J.A.: Statistical analysis of masked data. J. Official Stat. **9**(2), 407–426 (1993)

Little, R.J.A., Liu, F.: Selective multiple imputation of keys for statistical disclosure-control in microdata. Proceedings of the Section on Survey Research Methods, CD-ROM, American Statistical Association, Alexandria, VA 2002

Little, R.J.A., Liu, F.: Comparison of SMIKe with data-swapping and PRAM for statistical disclosure control of simulated microdata. Proceedings of the Section on Survey Research Methods, American Statistical Association, Seattle (2003)

Loeve, A.: Notes on sensitivity measures and protection levels. Technical report, Statistics Netherlands, The Hague (2001)

Machanavajjhala, A., Kifer, D., JAbowd, J., Gehrke, J., Vilhuber, L.: Privacy: from theory to practice on the map. ICDE Conference 2008. Cancun, Mexico, April 2008. http://www.cs.cornell.edu/johannes/papers/2008/icde2008-privacy.pdf (2008)

Mackey, E.: A framework for understanding statistical disclosure processes: a case study using the UK's neighbourhood statistics. PhD Thesis, Manchester University (2009)

Mackey, E., Elliot, M.J.: An application of game theory to understanding disclosure events. Proceedings of Work Session on Statistical Data Confidentiality, Bilbao, December 2009

Mackie, C., Bradburn, N.: Improving Access to and Confidentiality of Research Data. National Academy Press, Washington, DC (2000)

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., Walford, N.: The case for samples of anonymized records from the 1991 census. J. R. Stat. Soc. Ser. A **154**, 305–340 (1991)

Martin, D., Kifer, D., Machanavajjhala, A., Gehrke, J., Halpern, J.: Worst case background knowledge for privacy preserving data publishing. In Proceedings of the 23rd International conference on Data Engineering, ICDE 2007, April 15–20, Istanbul, Turkey (2007)

Mateo-Sanz, J.M., Domingo-Ferrer, J.: A method for data-oriented multivariate microaggregation. In: Domingo-Ferrer, J. (ed.) Statistical Data Protection, pp. 89–99. Office for Official Publications of the European Communities, Luxemburg (1999)

Mateo-Sanz, J.M., Sebé, F., Domingo-Ferrer, J.: Outlier protection in continuous microdata masking. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science, **vol. 3050**, pp. 201–215. Springer, Berlin, Heidelberg (2004)

McCullagh, K.: Data sensitivity: proposals for resolving the conundrum. J. Int. Commer. Law Technol. **2**(4), 190–201 (2007)

McCaa, R., Odinga, A.: Statistical confidentiality and the construction of anonymized public use census samples: a draft proposal for the Kenyan Microdata for 1989. Paper presented to Social Science History Annual Convention, Chicago, IL, 15–18 November 2001

McCaa, R., Ruggles, S.: The census in global perspective and the coming microdata revolution. Scand. Populat. Stud. **17**, 7–30 (2002)

McDonald, S.-K.: Recent experiences with resistance to national censuses in Western Europe. U.S. Bureau of the Census memorandum from Sarah-Kathryn McDonald, October 17, (1984)

McLuhan, M.: Understanding Media: The Extensions of Man. McGraw-Hill, New York, NY (1964)

McMillen, M.: Data access: national center for education statistics, of significance. J. Assoc. Public Data Users **2**, 1 (2001)

Mera, R.N.: Matrix masking methods which preserve moments. American Statistical Association Proceedings http://www.amstat.org/sections/srms/proceedings/papers/1997_075.pdf. Anaheim (1997)

Mokken, R.J., Kooiman, P., Pannekoek, J., Willenborg, L.C.R.J.: Disclosure risks for microdata. Stat. Neerl. **46**, 49–67 (1992)

Mollié, A.: Bayesian mapping of disease. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (eds.) Markov Chain Monte Carlo in Practice, pp. 359–379. Chapman and Hall, London (1996)

Moore, R.: Controlled data swapping techniques for masking public use data sets. U.S. Bureau of the Census, Statistical Research Division Report rr96/04. Available at http://www.census.gov/srd/papers/pdf/rr96-4.pdf. Accessed Jan 21, 2011 (1996)

Müller, W., Blien, U., Wirth, H.: Identification risks of micro data. Evidence from experimental studies. Sociol. Methods Res. **24**, 131–157 (1995)

Muralidhar, K., Parsa, R., Sarathy, R.: A general additive data perturbation method for database security. Manage. Sci. **45**(10), 1399–1415 (1999)

National Research Council: Expanding access to research data: reconciling risks and opportunities. Panel on Data Access for Research Purposes, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC (2005)

National Research Council: Putting people on the map: protecting confidentiality with linked social-spatial data. Panel on confidentiality issues arising from the integration of remotely sensed and self-identifying data. In: Gutmann, M.P., Stern, P.C. (eds.) Committee on the Human Dimensions of Global Change, Division of Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC. http://books.nap.edu/catalog.php?record_id=11865 (2007)

Navarro, A., Flores-Baez, L., Thompson, J.: Results of data switching simulation. Presented at the Spring meeting of the American Statistical Association and Population Statistics Census Advisory Committees, Washington, DC (1988)

National Academy Press: Expanding Access to Research Data: Reconciling Risks and Opportunities. http://www.nap.edu/openbook.php?record_id=11434&page=66 (2005)

Oganian, A., Domingo-Ferrer, J.: On the complexity of optimal microaggregation for statistical disclosure control. Stat. J. United Nations Econ. Commis. Eur. **18**, 4:345–354 (2001)

O'Keefe, C., Good, N.M.: A remote analysis server – what does regression output look like? In: Domingo-Ferrer, J., Saygýn, Y. (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science, **vol. 5262**, pp. 270–283. Springer, Berlin (2008)

Paass, G.: Disclosure risk and disclosure avoidance for microdata. J. Bus. Econ. Stat. **6**(4), 487–500 (1988)

Pagliuca, D., Seri, G.: Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey, Esprit SDC Project, Deliverable MI-3/D2 (1999)

Pickle, L.W., Waller, L.A., Lawson, A.B.: Current practices in cancer spatial data analysis: a call for guidance. Int. J. Health Geogr. **4**, 3 (2005)

Polettini, S.: Maximum entropy simulation for microdata protection. Stat. Comput. **13**(4), 307–320 (2003)

Polettini, S., Stander, J.: A Bayesian hierarchical model approach to risk estimation in statistical disclosure limitation. In Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases, pp. 247–261. Springer, Berlin (2004)

Purdam, K., Elliot, M.J.: An evaluation of the availability of public data sources which could be used for identification purposes – A Europe wide perspective, CASC project. University of Manchester, Manchester (2002)

Purdam, K., Elliot, M.J.: A case study of the impact of statistical disclosure control on data quality in the individual UK samples of anonymised records. Environ. Plann. A **39**, 1101–1118 (2007)

Purdam, K., Mackey, E., Elliot, M.J.: Whose data is it? Personal data and privacy. Paper presented to British Sociology Association Annual Conference, York 2003a

Purdam, K., Mackey, E., Elliot, M.J.: Personal data, privacy and the 2001 UK census. Paper presented to International Conference for the Development of the Information Society Conference, Lisbon, Portugal 2003b

Raghunathan, T.E.: Evaluation of inferences from multiple synthetic data sets created using semi-parametric approach. Panel on Confidential Data Access for Research Purposes, Committee on National Statistics, October (2003)

Raghunathan, T.E., Reiter, J.P., Rubin, D.R.: Multiple imputation for statistical disclosure limitation. J. Official Stat. **19**, 1–16 (2003)

Raghunathan, T.E., Rubin, D.R.: Multiple imputation for disclosure limitation. Department of Biostatistics Technical Report, University of Michigan, Dearborn, MI 2000

Rasinski, K.A., Wright, D.: Practical aspects of disclosure analysis, of significance: a. Topical J. Assoc. Public Data Users **2**(1), 35–41 (2000)

Rastogi, V., Suciu, D., Hong, S.: The boundary between privacy and utility in data publishing. Technical Report, University of Washington, Washington, DC 2007

Reiss, S.P.: Practical data-swapping: the first steps. ACM Trans. Database Syst. **9**, 20–37 (1984)

Reiss, S.P., Post, M.J., Dalenius, T.: Non-reversible privacy transformations. Proceedings of the ACM Symposium on Principles of Database Systems, Los Angeles, pp. 139–146 1982

Reiter, J.P.: Satisfying disclosure restrictions with synthetic data sets. J. Official Stat. **18**(4), 531–544 (2002)

Reiter, J.P.: Inference for partially synthetic, public use microdata sets. Surv. Methodol. **29**, 181–188 (2003)

Reiter, J.P.: Releasing multiply-imputed, synthetic public use microdata: an illustration and empirical study. J. R. Stat. Soc. Ser. A **168**, 185–205 (2005a)

Reiter, J.P.: Estimating risks of identification disclosure for microdata. J. Am. Stat. Assoc. **100**, 1103–1113 (2005b)

Reiter, J.P.: Releasing multiply imputed, synthetic public-use microdata: an illustration and empirical study. J. R. Stat. Soc. A **168**, 185–205 (2005c)

Reiter, J.P.: Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. J. Stat. Plann. Inference **131**, 365–377 (2005d)

"Restricted Access Procedures" by the Confidentiality and Data Access Committee (April 2002) at http://www.fcsm.gov/committees/cdac/cdacra9.doc

Robert, C.P., Casella, G.: Monte Carlo Statistical Methods (second edition). Springer, New York, NY (2004)

Robertson, D.A.: Cell suppression at statistics Canada. Proceedings of the Second International Conference on Statistical Confidentiality, Luxembourg 1994

Robertson, D.A.: Improving Statistics Canada's cell suppression software. Technical report, Statistics Canada, Ottawa, ON 2000

Robertson, D.A., Ethier, R.: Cell suppression: experience and theory. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases: From Theory to Practice. Lecture Notes in Computer Science, **vol. 2316**, pp. 8–20. Springer, New York, NY (2002)

Rubin, D.B.: Discussion of statistical disclosure limitation. J. Official Stat. **9**(2), 461–468 (1993)

Rubin, D.B.: Satisfying confidentiality constraints through the use of synthetic multiply-imputed microdata. J. Official Stat. **9**, 461–468 (1996)

Ruggles, S.: The public use microdata samples of the U.S. Census: research applications and privacy issues. A report of the Task Force on Census 2000, Minnesota Population Center and Inter-University Consortium for Political and Social Research Census 2000 Advisory Committee, Minnesota (2000)

Salazar, J.J.: A unified mathematical programming framework for different statistical disclosure limitation methods. Oper. Res. **53**(3), 819–829 (2005)

Salazar, J.J.: Controlled rounding and cell perturbation: statistical disclosure limitation methods for tabular data. Math. Program. **105**(2–3), 251–274 (2006a)

Salazar, J.J.: A new approach to round tabular data. In: Domingo-Ferrer, J., Franconi L. (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science, **vol. 4302**, pp. 25–34. Springer, Berlin, Heidelberg (2006b)

Salazar, J.J.: Statistical confidentiality: optimization techniques to protect tables. Comput. Oper. Res. **35**, 1638–1651 (2008)

Salazar, J.J.: Branch-and-cut versus cut-and-branch algorithms for cell suppression. In: Domingo-Ferrer, J., Magkos, E. (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science, **vol. 6344**, pp. 29–40. Springer, Berlin, Heidelberg (2010)

Salazar, J.J., Lowthian, P., Young, C., Merola, G., Bond, S., Brown, D.: Getting the best results in controlled rounding with the least effort. In: Domingo-Ferrer, J. (ed.) Privacy in Statistical Databases. Lecture Notes in Computer Science, **vol. 3050**, pp. 58–72. Springer, New York, NY (2004)

Samuels, S.J.: A Bayesian species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. J. Official Stat. **14**, 373–384 (1998)

Sande, G.: Automated cell suppression to preserve confidentiality of business statistics. Stat. J. United Nat. ECE **2**, 33–41 (1984)

Sande, G.: Structure of the ACS automated cell suppression system, In Statistical Data Confidentiality. Proceedings of the Joint Eurostat/UN-ECE Work Session on Statistical Confidentiality, Skopje, pp. 105–121 (1999)

Sande, G.: Exact and approximate methods for data directed microaggregation in one or more dimensions. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **10**(5), 459–476 (2002)

Sanil, A., Gomatam, S., Karr, A.: NISS WebSwap: a web-service for data swapping. J. Stat. Softw. 8(7) (2003). http://www.jstatsoft.org/v08/i07

Schlörer, J.: Security of statistical databases: multidimensional transformation. ACM Trans. Database Syst. **6**(1), 95–112 (1981)

Shlomo, N.: Accessing microdata via the internet. Joint UN/ECE and Eurostat Work Session on Statistical Data Confidentiality, Working Paper No. 6. Luxembourg, April 7–9 (2003)

Sieber, J.E.: Privacy and Confidentiality: As Related to Human Research in Social and Behavioral Science (Research Involving Human Participants V2): Online Ethics Center for Engineering, May 25, National Academy of Engineering www.onlineethics.org/CMS/research/resref/nbacindex/nbachindex/hsieber.aspx. Accessed Sept 25, 2007 (2007)

Simard, M.: Development of a real time remote access infrastructure at statistics Canada. Paper presented at the joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Bilbao, Spain, 2–4 December 2009

Singer, E., Mathiowetz, N.A., Couper, M.P.: The impact of privacy and confidentiality concerns on survey participation: the case of the 1990 U.S Census. Public Opin. Q. **57**(4), (Winter 1993), 465–482 (1993)

Singer, E., Van Hoewyk, J., Neugebauer, R.J.: Attitudes and behavior – the impact of privacy and confidentiality concerns on participation in the 2000 census. Public Opin. Q. **67**, 368–384 (2003)

Skinner, C.J.: Statistical disclosure issues for census microdata. Paper presented at International Symposium on Statistical Disclosure Avoidance, Voorburg, The Netherlands, 13 December 1990

Skinner, C.J., Elliot, M.J.: A measure of disclosure risk for microdata. J. R. Stat. Soc. Ser. B **64**(4), 855–867 (2002)

Skinner, C.J., Holmes, D.J.: Modelling population uniqueness. Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute, Luxembourg, pp. 175–199 (1992)

Skinner, C.J., Shlomo, N.: Assessing identification risk in survey micro-data. J. Am. Stat. Assoc. **103**(483), 989–1001 (2008)

Skinner, C.J., Holmes, D.J.: Estimating the re-identification risk per record in microdata. J. Off. Stat. **14**, 361–372 (1998)

Sparks, R., Carter, C., Donnelly, J., O'Keefe, C.M., Duncan, J., Keighley, T., McAullay, D.: Remote access methods for exploratory data analysis and statistical modelling: privacy-preserving analytics TM. Comput. Methods Progr. Biomed. Arch. **91**(3), 208–222 (2008)

Smith, D., Elliot, M.J.: A measure of disclosure risk for aggregate data. Paper presented to the International Conference of the Royal Statistics Society, Turin, September 2004

Smith, D., Elliot, M.J.: An experiment in Naive Bayesian record linkage. Proceedings of Conference of the International Statistical Institute, Sydney, April 2005

Smith, D., Elliot, M.J.: A measure of disclosure risk for tables of counts. Trans. Data Priv. **1**(1), 34–52 With Smith, D. (2008)

Spruill, N.L.: The confidentiality and analytic usefulness of masked business microdata. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 602–607. Alexandria, VA (1983)

Statistics Netherlands: Argus User's Manual. http://neon.vb.cbs.nl/casc/ (2007)

Steel, P., Sperling, J.: The Impact of Multiple Geographies and Geographic Detail on Disclosure Risk: Interactions between Census Tract and ZIP Code Tabulation Geography. U.S. Census Bureau, Washington (2001)

Stigler, S.M.: Adolphe Quetelet. Encyclopedia of Statistical Sciences. Wiley, New York, NY (1986)

Sullivan, G., Fuller, W.A.: The use of measurement error to avoid disclosure. Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 802–807 (1989)

Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **10**(5), 571–588 (2002)

Takemura, A.: Local recoding and record swapping by maximum weight matching for disclosure control of microdata sets. J. Official Stat. **18**(2), 275–289 (2002)

Tendick, P.: Optimal noise addition for preserving confidentiality in multivariate data. J. Stat. Plann. Inference **27**, 341–353 (1991)

Tendick, P., Matloff, N.: A modified random perturbation method for database security. ACM Trans. Database Syst. **19**, 47–63 (1994)

Thibaudeau, Y., Winkler, W.E.: Bayesian networks representations, generalized imputation, and synthetic microdata satisfying analytic restraints. Statistical Research Division Report RR 2002/09, Washington. http://www.census.gov/srd/www/byyear.html (2002)

Torra, V.: Microaggregation for categorical variables: a median based approach. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science, **vol. 3050**, pp. 162–174. Springer, Berlin, Heidelberg (2004)

Torra, V., Abowd, J., Domingo-Ferrer, J.: Using Mahalanobis distance-based record linkage for disclosure risk assessment. Lect. Notes in Comput. Sci. **4302**, 233–242 (2006)

Tranmer, M., Pickles, A., Fieldhouse, E., Elliot, M.J., Dale, A., Brown, M., Martin, D., Steel, D., Gardiner, C.: Microdata for small areas. J. R. Stat. Soc. Ser. A **168**(1), 29–49 (2005)

Tranmer, M., Steel, D., Chambers, R., Clark, R., Elliot, M.: The role of individuals, geographical groups, households and social networks in social statistics. Paper presented to the 30th Sunbelt Conference Riva del Garda, Italy, June 2010

Trottini, M.: A decision-theoretic approach to data disclosure problems. Paper prepared for 2nd Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje, Macedonia, 14–16 March 2001 (2001)

Trottini, M.: Decision models for data disclosure limitation. Ph.D. thesis, Department of Statistics, Carnegie Mellon University (2003)

Trottini, M.: Statistical disclosure limitation in longitudinal linked data: objectives and attributes. UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva, Monograph in Official Statistics, Eurostat, 165–173 (2005)

United Nations: Managing Statistical Confidentiality & Microdata Access: Principles and Guidelines of Good Practice (2007)

Vaidya, J., Clifton, C., Zhu, M.: Privacy Preserving Data Mining. Springer, Heidelberg (2006)

Valls, V., Torra, V., Domingo-Ferrer, J.: Aggregation methods to evaluate multiple protected versions of the same confidential data set. In: Grzegorzewski, P., Hryniewicz, O., Gil, M.A. (eds.) Soft Methods in Probability, Statistics and Data Analysis, pp. 355–362. Series Advances in Soft Computing. Physica, Heidelberg (2002)

Willenborg, L.C.R., de Waal, T.: Statistical Disclosure Control in Practice. Lecture Notes in Statistics, **vol. 111**. Springer, New York, NY (1996)

Willenborg, L.C.R., de Waal, T.: Elements of Statistical Disclosure Control. Lecture Notes in Statistics, **vol. 155**. Springer, New York, NY (2001)

Winkler, W.E.: Matching and record linkage. Technical report RR93/08, Statistical Research Division, U.S. Bureau of the Census, Washington, DC 1993

Winkler, W.E.: Advanced methods for record linkage. Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 467–472 1994

Winkler, W.E.: Matching and record linkage. In: Cox, B.G. et al. (ed.) Business Survey Methods, pp. 355–384. Wiley, New York, NY (1995a)

Winkler, W.E.: Advanced methods for record linkage. Proceedings of the American Statistical Association Section on Survey Research Methods, Alexandria, VA, pp. 467–472 1995b

Winkler, W.E.: Re-identification methods for evaluating the confidentiality of analytically valid microdata. Res. Official Stat. **1**, 87–104 (1998)

Winkler, W.E.: Masking and re-identification methods for public-use microdata: overview and research problems. RESEARCH REPORT SERIES (Statistics #2004-06): Statistical Research Division, U.S. Bureau of the Census http://www.census.gov/srd/papers/pdf/rrs2004-06.pdf (2004). Accessed Oct 5, 2007

Winkler, W.E.: Masking and re-identification methods for public-use microdata: overview and research problems. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science, **vol. 3050**, pp. 231–246. Springer, Berlin, Heidelberg. http://www.census.gov/srd/papers/pdf/rrs2004-06.pdf (2004a)

Winkler, W.E.: Re-identification methods for masked microdata. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases, pp. 216–230. Springer, New York, NY (2004b)

Wu, J., Abowd, J.: Synthetic data for administrative record applications at LEHD. Available online in the LEHD Presentations Library (2008)

Xiao, X., Wang, G., Gehrke, J.: Interactive anonymization of sensitive data. SIGMOD'09 June 29–July 2, 2009, Providence, Rhode Island, USA. ACM 978-1-60558-551-2/09/06 (2009)

Yancey, W.E., Winkler, W.E., Creecy, R.H.: Disclosure risk assessment in perturbative microdata protection. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. Lecture Notes in Computer Science, **vol. 2316**, pp. 135–152. Springer, Berlin, Heidelberg (2002)

Zaslavsky, A.M., Horton, N.J.: Balancing disclosure risk against the loss of nonpublication. J. Official Stat. **14**, 411–419 (1998)

Zayatz, L.V.: Estimation of the percent of unique population elements in a microdata file using the sample. Statistical Research Division Report Series, Census/SRD/RR-91/08 Bureau of the Census, Washington, DC, pp. 674–684 1991

Zayatz, L.V.: Using linear programming methodology for disclosure avoidance purposes, Technical report, U.S. Bureau of the Census. Research Report RR-92/02, Washington, DC 1992

Zayatz, L.V., Massell, P., Steel, P.: Disclosure limitation practices and research at the U.S. Census Bureau. Special issue on statistical disclosure control. Neth. Off. Stat. **14**, 26–29 (1999)

Zayatz, L.V., Rowland, S.: Disclosure limitation for American FactFinder. Paper presented at the American Statistical Association Joint Statistical Meetings, Baltimore, MD, 8 August 1999

Index