Chapter 1
Introduction*

1. The Trade-off Dilemma

There is a fundamental tension at the heart of every statistical agency's mission. Each is charged with collecting high quality data to inform national policy and enable statistical research. This goal necessitates dissemination of both summary data and microdata. Each is also charged with protecting the confidentiality of survey respondents—not only because of legal and ethical mandates, but because public trust and perceptions of that trust are important contributors to data quality and response rates.

Protecting confidentiality necessitates perturbing or summarizing the data in some fashion so that the individual respondent cannot be identified. Greater protection of confidentiality means that the data, which cost so much to collect and produce, are likely to become less valuable. The resulting trade-off dilemma, which could well be stated as protecting confidentiality (avoiding disclosure) but maximizing data quality and data access, has become more complex as both technological advances and public perceptions have changed in this Information Age. In sum, while statistical agencies go to great lengths to collect high quality data, the necessity of protecting confidentiality results in some data quality compromises. This book describes new theoretical, practical, and technological responses to the challenges that statistical agencies face.

What are these challenges, and how have they changed the world within which agencies collect and protect data? A partial list would include an increasing demand by policy makers for timely, relevant information; an increasing demand by academic researchers for microdata; the public dissemination of state- and local-level administrative records; and possibly most important, the increasing data collection by the private sector.

* The editors would like to thank the following reviewers: Nancy Bates, Larry Cahoon, Cynthia Clark, James Fagan, Gerald Gates, Nancy Gordon, Jennifer Guarino, Joan Marie Hill, Frederick Knickerbocker, Larry Long, Paul Massell, Randall Neugebauer, Carole Popoff, Juanita Rasmann, Arnie Reznek, Stephanie Shipp, Phil Steel, Sam Hawala, Judith Waldrop, Diane Willimack, and Tommy Wright. We also thank Felicity Skidmore, for excellent editing assistance, and Jayne Sutton of EEI Communications for her technical expertise in preparing the final manuscript.

Information is critically important to policy makers—it permits the management of the economy and it informs public debate. Data ranging from inflation rates to unemployment rates, crime statistics, and healthcare statistics are all important components of our information infrastructure. As the economy becomes increasingly more complex, however, and the interactions among households, businesses, and governments more entangled, the demand for data describing these interactions has increased. At the same time, the cost and burden on respondents has increased, so statistical agencies have turned more often to administrative data to respond to user needs. Some state and local governments provide drivers' license records and property ownership information on-line. The Netherlands no longer conducts traditional population censuses; instead, it relies on administrative records to count the size of the population. Although these uses provide high value added from existing data without imposing additional burden on respondents, they do pose special confidentiality challenges. This book addresses some of the solutions to these challenges.

External researchers increasingly demand access to detailed data at the micro level for two reasons. First, no microdata on business issues are available to the public from statistical agencies, because these data are protected from disclosure, yet important new statistics on job creation and job destruction can only be derived from such data. Second, increasingly stringent disclosure protection has meant that publicly available data on households and individuals are increasingly too geographically aggregated, or income levels too strictly top-coded, to allow full policy analysis.

Statistical agencies have an incentive to meet the needs of researchers not only because it fulfills their core mission of data dissemination, but also because research access can often lead to the core improvement of statistical data products. Analytically sophisticated projects, mounted by external researchers, can be used to assess whether the data produced by statistical agencies are of sufficient quality for policy makers to act on the results. In this sense, giving external researchers access to data can act to complement the standard data quality control checks used by statistical agency staff.

In short, the process of creating analytical results (and hence, if the research is successful, informing policy makers) is integral to evaluating and improving the data that statistical agencies produce—quite apart from the scientific value contributed by the analysis of the data themselves. As the economy becomes more complex and the questions posed by policy makers more detailed, there is an increasing need for external researchers to have access to micro-level data without violating disclosure guidelines. Again, some chapters in this book explain how this can be done.

Data collection by the private sector has soared. This phenomenon is likely to interact in complex ways with both the ability of statistical agencies to protect the data they collect and the public perception of that ability. On the one hand, the vast expansion of private data collection creates a much more sophisticated master file

for potential intruders. Latanya Sweeney (Chapter 3) gives a sense of just how vast private collections are in her example of Catalina Marketing. This company began in March 1996 to store the shopping patterns of 143 million shoppers each week from 11,000 supermarkets nationwide. By July 1998, it had amassed a 2-terabyte database with 18 billion rows of data. Widespread awareness of these types of data collections may make the public more concerned by the increased potential for re-identification. On the other hand, because statistical agencies are much more heavily regulated than the private sector and heavy punishments apply for disclosure, the public may be reassured. Chapters in this book inform us of both the ways in which statistical agencies have responded to the technological challenges and the effect their response has had on public perceptions.

How have statistical agencies adapted to these challenges? The historical approach has been to protect data in three ways: statistical analysis, data protection, and different access modalities. Statistical analysis protects confidentiality because information about each respondent is used only in the context of information collected from other respondents. That is, information is not obtained based on the experience of just one or two individuals, but rather on trends observed in groups of individuals, and the safety in numbers approach is at work to protect individual respondents.

Data protection occurs when data producers take steps to modify or suppress information that might identify an individual either directly or indirectly before they make it available to the public for analysis. This process, which is also called disclosure protection, has evolved from the simple suppression of information on unique observations to complex statistical methods affecting a large amount of the information collected. The growth in the level of sophistication of data protection mirrors the expansion of technology that could potentially permit people to circumvent the protection process. Some of the most technical chapters in the book deal directly with new advances in this area.

New access modalities have also evolved—from licensing, to remote access, to secure remote sites. The book has chapters on each of these topics.

How does the public perceive the way in which statistical agencies respond to these challenges? Perceptions matter, both because statistical agencies need to maintain response rates and data quality and because they need to reassure the public that they are fulfilling their legal and ethical requirements. If the public has become more sensitive to data privacy concerns in response to private sector actions, then statistical agencies need to respond accordingly.

The assessment of public perceptions can also be an important tool in guiding data dissemination decisions. For example, new access modalities such as restricted access sites are potentially important new dissemination mechanisms. Their establishment should be evaluated, however, not only for the technical protection and the resulting data quality issues but also in the context of public perception of these approaches. It may well be that the public perceives them as being much safer than the release of public use files, particularly given the wealth of pri-

vate data available, and statistical institutes can tailor their response appropriately. Three of the chapters in this book explicitly address the perception issues, for businesses and for households.

This book provides an overview of how statistical agencies have risen to the new challenges confronting them and could rise to future challenges. It begins with two chapters that delineate the new technological challenges faced and summarize the types of approaches that different statistical agencies use to protect data. The subsequent set of chapters review and develop new state-of-the-art techniques that directly address these statistical disclosure techniques from both theoretical and practical perspectives. The next section's chapters describe alternative access modalities. The book concludes with chapters that update our knowledge on perception issues.

2. Fundamental Concepts

In any technical monograph like this one, the material is most efficiently presented if basic technical terms are used. So that all readers are able to appreciate the material in this book, we provide a discussion of underlying concepts and some technical definitions. The brochure *Confidentiality and Data Access Issues Among Federal Agencies* is the source of much of the material that follows. The authors thank the publisher of the brochure, the Confidentiality and Data Access Committee, sponsored by the Federal Committee on Statistical Methodology.

Data Structures

Data and thus disclosure methods come in two basic forms: microdata that refer to individual units and aggregate estimates from survey or census responses. Two data formats commonly used to present aggregate estimates are tables of frequency counts and tables of aggregate magnitude data.

Microdata files consist of individual records that contain values of variables for a single person, a business establishment, or another individual unit. Public use microdata files are released to the public for research and analytical purposes after being subjected to procedures that limit the risk of disclosure.

Frequency count tables count the number of respondents with specified characteristics. For example, a two-dimensional frequency count table may have rows corresponding to race categories and columns corresponding to age groupings. An individual table cell at the intersection of a given row and a given column would indicate the number of residents of a certain geographic area with that race and age combination.

Tables of *aggregate magnitude data* are analogous to frequency count tables in that they are defined by cross-classification of categorical variables. However, the cells contain aggregate values, over the corresponding respondents, of some quan-

tity of interest. For example, a two-dimensional table on income defined by race and age would contain total incomes in each race-by-age cell.

A special case of tabular data are systems that permit generation of tables on demand through tabulations from an on-line query to a statistical database using the Internet or other forms of remote access. Data users may create their own tabulations by customized queries to the database.

Disclosure Limitation Methods

Different approaches to disclosure protection apply depending on the type of underlying data to be protected.

In the case of microdata, three common types of procedures are applied to prevent disclosure of confidential information. First, information that directly reveals the identity of the respondents is suppressed. Second, information that may indirectly reveal the identity of a respondent is suppressed. This can be accomplished by reducing the variation within the data through rounding, top- and bottom-coding, collapsing response categories, and suppressing information such as detailed geography. Third, some uncertainty can be introduced into the reported data. This can be accomplished by altering the underlying data through swapping of reported values among similar respondents, adding predetermined random noise to the data, and performing other more structured randomization of the data.

In the case of aggregate magnitude data and some frequency count data, some information is suppressed. *Primary cell suppression* is withholding information in a cell because its publication would explicitly or implicitly reveal confidential or sensitive information. When a table contains cells that represent sums of either rows or columns, primary cell suppression alone does not always protect the confidential data. Original values of primary cells may be determined exactly or within a narrow range through subtraction. When this occurs, it is necessary to perform *complementary cell suppression* to protect the primary suppressions from disclosure. In addition to cell suppression, agencies may use rounding, data swapping, or some other type of noise addition to protect frequency count data.

Systems that support custom tabulations of sensitive data also provide disclosure protection so that all data extracted directly from these systems are adequate for public dissemination. An issue unique to this form of data access is that of complementary disclosure. If two or more separate requests are combined outside the system, it might create a disclosure problem that did not exist for the individual requests.

Aside from table-on-demand systems, data producers can provide the opportunity for outside researchers to create custom tables and models from non-public data through *licensing, secure sites,* and *remote access.* Licensing is an arrangement whereby an institution and its researchers sign formal agreements to protect the data according to the laws and policies governing the data collection agencies. In return they can receive the data for a limited period of time. Secure sites repre-

sent offices under the data production agencies' control that are made available to researchers who formally agree to follow the laws and policies of the data collection agencies. They also agree to subject their results to disclosure review before the results are removed from the offices. Remote access is an arrangement whereby researchers provide computer programs to the data collection agencies, which execute them and review the results for disclosure violations before sending them to the researcher.

3. The Contribution of the Book

Overview

The book starts with two overview chapters that summarize current practical approaches in key North American and European countries to protect confidentiality while maximizing access to economic and demographic data. They also present the current challenges in protecting the confidentiality of data on individuals that arises from the ever-increasing volume of information available from public and private sources and the ever-increasing sophistication of technology in accessing and linking this information.

Flóra Felsö, Jules Theeuwes, and Gert. G. Wagner (Chapter 2) conducted a survey of several national statistical offices on disclosure control methods in use, receiving responses from Canada, the Czech Republic, Denmark, Estonia, France, Germany, Hungary, Italy, Lithuania, the Netherlands, New Zealand, Norway, Sweden, and four agencies in the United States. The authors begin their chapter with a literature review focusing mainly on *Statistical Policy Working Papers #2* (Federal Committee on Statistical Methodology 1978) and *#22* (Federal Committee on Statistical Methodology 1994). These papers discuss disclosure limitation procedures for various types of data, including microdata, frequency count data, and aggregate magnitude data from both demographic and establishment surveys and censuses. The survey questions were based on *Statistical Policy Working Paper #22* and again covered all types of data.

In summarizing the survey results, the authors found that most agencies use a threshold rule (quite often the threshold is 3) for determining which cells in a frequency count table are sensitive (potential disclosure problems), and these agencies typically use cell suppression to protect such cells. For aggregate magnitude data, most agencies use a technique called the N-K rule to determine which cells are sensitive and use cell suppression and/or table reconstruction to protect them. Many agencies do not release microdata to the public. Those that do eliminate obvious identifiers, limit geographic detail, and limit the number and detail of variables. There was no consensus on what software packages should be used for purposes of limiting disclosure.

The explosion of private data collection is documented by Latanya Sweeney (Chapter 3). She describes three approaches to data collection in the private data sector today: collect more, collect specifically, and collect it if you can. The implications of this behavior in terms of the potential to identify respondents are troubling because they potentially lead to many more match keys that could be used to link publicly funded masked microdata to privately funded or administrative databases with identifiers. For example, historically, birth certificate information had only 7 to 15 fields of information, but today more than 100 fields of information are collected about each child's birth. This situation clearly poses a threat to statistical agencies because they could potentially be used in conjunction with statistical data products in attempts to uncover a survey or census respondent's information. Sweeney speculates that past practices may no longer be applicable guides because of the amount of data now being collected and current technology. Based on an examination of some information released by the federal government that was not subject to strict disclosure controls, she also speculates that 'current policies and practices support crude decisions'. Fortunately, Felsö, Theeuwes, and Wagner (Chapter 2) demonstrate that most efforts to protect confidentiality are more stringent than the ones Sweeney has studied, so that her conjecture does not apply across the board.

Public Access Through Data Manipulation

The second group of chapters focuses on the problem of providing public access to high quality information to inform public debate and research without disclosing information on individuals or businesses. This approach is one of data manipulation to mask the underlying information in the data to avoid disclosure while minimizing the impact of that manipulation on estimates derived from the public data. These chapters

- Describe different approaches to measuring the disclosure risk associated with uses of data on individuals and households and test the success of those methods in preventing disclosure.
- Discuss disclosure methods used to protect information included in microdata and the impact of these disclosure methods on the estimates produced from the microdata (known as information loss).
- Summarize disclosure methods for frequency count data, identify methods for measuring information loss, and quantify the loss under various disclosure limitation techniques.
- Summarize exploratory research to develop connections between current and proposed selected disclosure rules and disclosure limitation methods for economic tabular data and the secondary cell suppression problem as a mathematical statement of the problem of avoiding disclosure but maximizing access for tabular data.

- Present methods used to assess disclosure risk for aggregate data. The chapters address issues such as the appropriate primary suppression rules for tabular data to avoid disclosing a particular institution's response and approaches to reduce disclosure risk for economic tabular data, information loss resulting from cell suppression and/or recoding, and secondary cell suppression software.
- Discuss the use of linked data; that is, datasets created through matching of information from two or more sources such as data on employers from establishment sources and data on employees from surveys of individuals. The authors pay particular attention to the public release of parameters of models estimated using data that cannot be released for disclosure reasons.

In understanding disclosure risk, it is important to decompose the different factors that contribute to that risk. Mark Elliot (Chapter 4) does this, focusing on disclosure risk for public use microdata files, which are typically produced from demographic surveys or censuses. He discusses an 11-point disclosure attempt scenario in which he decomposes the risk into, for example, the motivation for an intruder to make such an attempt, the means necessary to carry out the intrusion, and the effect of data divergence between the intruder's information and the target file. In so doing, he identifies the critical elements that contribute to the risk of disclosure for each given file: the sampling fraction, level of detail on variables, level of geographic detail, number of key variables that might be used for linking purposes, and data divergence. He describes two reidentification studies that indicate the difficulties faced by an intruder identifying an individual in demographic microdata—in particular, the large number of false positive matches. He also develops a data intrusion simulation that could be used in an attempt to measure the disclosure risk of a microdata file, noting that disclosure risk assessment for microdata is still a young and complex research area.

We noted earlier that statistical agencies face a fundamental trade-off between data quality and data protection. If decisions must be made about data protection, it would be enormously helpful to have some measure of this trade-off. Josep Domingo-Ferrer and Vicenç Torra (Chapters 5 and 6) provide an overview of the disclosure limitation methods for both continuous and categorical variables in microdata. Several types of available methods are described, and the authors measure the effect of each method on the disclosure risk of the data and the resulting information loss. They measure disclosure risk using distance-based record linkage (linking the original data to the masked data), probabilistic record linkage, and interval disclosure. They measure information loss for continuous data in terms of mean square error, mean absolute error, and mean variation between covariance, correlation, and other matrices used for data analysis. They measure information loss for categorical data in terms of a comparison of contingency tables and entropy-based measures. The methods tested for continuous data were additive noise, data distortion by probability distribution, resampling, micro aggregation, lossy compression, and rank swapping. Of those methods, rank swapping seemed to perform best. The methods tested for categorical data were top- and bottom-coding,

global recoding, and post-randomization (PRAM). Of these methods, top- and bottom-coding and global recoding all performed well, while PRAM did not.

George T. Duncan and colleagues (Chapter 7) provide an equally interesting framework for disclosure risk and information loss for frequency count data via a graph they call an R-U confidentiality map. They present methods of auditing tables and sets of related tables[1] to look for potential disclosure problems. These methods include linear and integer programming, generalizations of Frechet and Bonferroni bounds, and a generalization of Buzzigoli and Giusti's shuttle algorithm. The authors describe methods for protecting frequency count data, including sampling, cell suppression, local suppression, rounding, data swapping, simulated tables, and Markov perturbation, and they note advantages and disadvantages of each. They compare cell suppression, rounding, and Markov perturbation with different parameters via the confidentiality map and show that determining the best procedure may depend on the level of risk an agency finds acceptable.

While the previous set of chapters looked at the disclosure issues associated with demographic data, Lawrence H. Cox (Chapter 8) examines disclosure risk in aggregate magnitude data in the context of economic surveys and censuses. This disclosure poses special problems because large firms are very easily identified if industry or geographic detail is provided (the classic example is General Motors in Detroit, but one could just as easily think of Microsoft in Washington State). Cell suppression is the method typically used to protect this type of data. The author discusses the typical structure of the data and how sensitive cells (primary suppressions) are identified via different ways of quantifying risk. He then describes complementary suppression and various ways to calculate information loss, including number of cells suppressed and total value suppressed. Until recently, complementary cell suppression was done for one primary suppression at a time. Fischetti and Salazar have developed and tested a method for performing complementary cell suppression for all primary suppressions simultaneously. This method can reduce oversuppression and allow for more data to be published. Their algorithm has some drawbacks, the largest of which is that it protects data at the establishment level while most agencies must protect data at the company level. Cox has developed a similar algorithm designed to protect data at the company level. This algorithm needs to be examined and improved upon based on computational considerations.

While Cox provides a very theoretical approach to the core disclosure problems posed by economic data, Sarah Giessing (Chapter 9) provides a description from a practitioner's point of view. Her chapter provides numerous examples with different ways to quantify risks and differences between various approaches—particularly discussing minimum size requirements for complementary suppressions, one cell's capacity for protecting another, cell suppression patterns, and protection intervals around suppressed cells. The author discusses

[1] This is particularly important because often individual tables do not reveal individual identities, but can present problems when examined in conjunction with one another.

heuristic approaches to complementary cell suppression such as the hypercube approach, the network flow approach, the linear programming approach, and the integer linear programming approach. She then mentions the currently available cell suppression software systems and highlights some key attributes of that software, such as computing time and resource requirements, data structure and software implementation, the ability to process linked tables, and the ability to assign preferences to choose or not choose certain cells as complementary suppressions.

John Abowd and Simon Woodcock (Chapter 10) present methods for disclosure limitation of longitudinal linked data. Longitudinal linked microdata contain observations from two or more related sampling frames with measurements for multiple time periods from all units of observation. The prototypical longitudinal linked data set that they consider contains observations about individuals, work histories, and employers. They present methods for disclosure limitation of parameter estimates obtained from analyses of such data, as well as conditional expectations such as contingency tables and summary statistics. They also present a method for disclosure limitation of the microdata itself that is based on multiple imputation techniques developed for missing data. Disclosure limitation of longitudinal linked data is complex because of the requirement that new data be disclosure-proofed in a manner consistent with both the underlying microdata and previous disclosure-proofed releases. In the particular application they consider, this complexity is intensified by the fundamental differences in the statistical properties of data on individuals and data on businesses.

Remote Access to Non-Public Data

All the techniques described thus far necessarily involve data manipulation or suppression and are likely to reduce the quality of estimates to be produced from data sources. As a result, statistical agencies have begun to investigate other methods that allow use of data while protecting confidentiality of the respondents. These methods allow the data to be used in an environment controlled by the data-producing agency and require that its use be subject to the same legal and ethical protections placed on the agency itself. This group of chapters

- Introduces the process of licensing whereby institutions and researchers outside the data-producing agencies temporarily gain access to data at their site by agreeing to conform to the legal protections surrounding those data that are imposed on the data-producing agency.
- Describes secure sites, where the data remain under the control of the data-producing agencies and researchers come to an agency office to access it. Such sites are an increasingly popular means of providing researchers access to respondent-level microdata while protecting the confidentiality of the data.

- Demonstrates an approach to house the data at the data-producing agency and allow remote access by researchers through an intermediary controlled by the agency that guarantees all use conforms to the law.

Data licensing is a way to provide access to data when they cannot be released to the public because of confidentiality concerns. It is described in some detail by Marilyn M. Seastrom (Chapter 11). A number of U.S. statistical agencies currently use licensing, but she focuses on the licensing system at the National Center for Education Statistics. She describes in detail the license application, required security procedures, who can access the licensed data, publishing requirements, security inspections, and the termination of licenses. The author discusses various U.S. laws and regulations that agencies use in licensing and then compares how various agencies implement and enforce licensing agreements. She gives examples of both major and minor violations of licensing agreements. She concludes by recommending that all agencies that license data perform periodic inspections of the licensed sites and develop and maintain a database application that allows the agency to readily access records of licensed files and authorized users for each agreement.

Probably the most important access modality developed in the past decade is that of restricted access sites. These sites permit statistical agencies to respond to the microdata needs of researchers, avoid the linkage problems posed by the Internet, and address potential perception problems that might be associated with other access modalities. Timothy Dunne (Chapter 12) discusses the establishment and management of such secure research sites to provide access to data when they cannot be released to the public because of confidentiality concerns. He notes that an agency must first decide if it legally has the authority to establish such a site and under what conditions the data may be accessed. The next steps, described in detail, are to choose the physical location and establish security and personnel there. The agency must then focus on what data will be available at the secure site and how they will be managed. Dunne then addresses project management issues, including project selection, formal agreements that must be established between the researcher and the agency, researcher training, and reviewing results for potential disclosure problems. If the secure site is at a non-agency location, the agency must consider how a site is awarded and how it will be managed. The author highlights many benefits of establishing secure sites, including the development of a community of skilled data analysts, the development of new data and statistical products, analysis of longitudinal and/or linked data, and feedback to the agency on methodological aspects and data quality issues of the data being analyzed.

Michael Blakemore (Chapter 13) presents the potentials and the perils of remote access. He stresses how rapidly developing information technologies offer increasing potential for unrestricted information flow. At the same time, he notes, the development of information technology increases the costs to data custodians of making mistakes, one cost being a potential loss of trust on the part of data providers. Remote access has three key components: the network, along with the physical

information technology infrastructure; the software, ranging from security systems to encryption; and finally, the organization context. These key elements are highlighted. The chapter concludes with a set of case studies on ways in which remote access has been implemented.

Perceptions

Regardless of the extent and success of the measures used to protect individual information, some people still believe their data cannot be protected, and this perception could have a detrimental impact on their participation in surveys. This series of chapters

- Summarizes the research on the public's attitudes and perceptions toward privacy and confidentiality.
- Addresses how individuals' beliefs about disclosure of personal information are influenced by historic mistrust of government in groups considered hard to enumerate.
- Discusses perceptions among businesses, organizations, and institutions, contributing to the growing body of literature in the demographic sector on the effect of the perceptions of protection and perceptions of harm on respondents' willingness to participate in surveys.

The chapter by Eleanor Singer (Chapter 14) examines in some detail what the public believes about the confidentiality of data collected by statistical agencies and how it regards the prospect of data sharing among federal agencies. Singer also examines changes in the public's beliefs and attitudes over time and how these beliefs may affect response rates to demographic surveys and censuses. The chapter is based primarily on four surveys undertaken (in 1995, 1996, 1999, and 2000) by the U.S. Census Bureau that tracked attitudes about privacy and confidentiality, primarily in relation to the census of population and housing. The author describes in detail the methods used, trends in beliefs about confidentiality and attitudes toward privacy, trends in attitudes toward sharing of data among federal statistical agencies, predictors of privacy-related attitudes, and the relationship between attitudes and behavior. Looking across the four surveys and from the perspective of five years, one can see distinct patterns of change with respect to knowledge and awareness of the census itself and with respect to knowledge about Census Bureau confidentiality practices. There is a secular increase in knowledge about confidentiality, which is paralleled by a significant increase over time in the percentage of respondents who would be bothered if their census data were provided to anyone outside the Census Bureau. Interestingly enough, these changes are *not* paralleled by increasing distrust of data usage or increasing concerns about privacy or by declining trust in the government.

What affects public response to requests for information in government surveys? Eleanor R. Gerber summarizes research on this topic (Chapter 15). The core focus is on modeling respondents' decisions to provide (or refuse to provide) infor-

mation: how this decision is made, what factors are taken into account, and what other concerns are evoked in considering this decision. Exploratory qualitative techniques (ethnography) rather than numerical assessments were used for this work. The author finds that people like to feel that they are in control of information about them. This feeling affects their attitudes toward data sharing and toward different modes of questionnaire administration. Perception of a legitimate need for the information (including benefiting society as a whole) is a critical factor in whether a respondent decides to provide that information. These data strongly suggest that trust (or lack thereof) in assurances of confidentiality is only one element in a complex set of attitudes toward privacy in general. The public perception is that data are widely exchanged among government agencies, and many people view this exchange as a loss of control of their information. The author recommends stressing the legitimate need for the information being collected and addressing data use concerns at the time of data collection.

Although a great deal of research has been targeted at understanding people's views of confidentiality, very little has addressed business perceptions of confidentiality protections. Nick Greenia and colleagues (Chapter 16) describe the results of one of the first surveys conducted in this area. The survey focused on the sensitivity of the individual data items, perceived benefits of the data collection, cost of the data collection, and the protection provided to the respondents. The results of this survey show that a wide variety of businesses distrust the government, but a large number of businesses would actually be amenable to some sharing of data among agencies and the release of older and less sensitive business data to the general public.

4. Implications: 'There's No Data Like No Data'

The trade-off dilemma mentioned in the opening paragraphs of this introductory chapter is central to this book. Statistical agencies want to protect the confidentiality of survey respondents and avoid disclosure while at the same time maximizing data quality and data access. While the fundamental, hard-to-solve tension between these two objectives has always existed, the tension is exacerbated by the increased ability of modern society to generate more information and the expanded desire for fast and accurate information on complex societal problems. Finding solutions to the confidentiality problems thus posed should be paramount because the natural tendency of statistical agencies, when faced with uncertainty about the impact of the release of information, is to err on the side of protecting confidentiality. The concomitant risk, which is the reduced quality or quantity of publicly available data, is a very real one without alternative legal approaches to permit access to data for research directed toward public policy issues.

This book addresses some issues associated with the core trade-off dilemma. In particular, it provides an up-to-date overview of tools that are available for extend-

ing access to data users. It discusses the wide array of instruments that are available to statistical agencies as they seek resolution of their trade-off dilemma. These instruments range from use of disclosure limitation procedures when releasing data to providing remote access through data licensing and restricted access sites. It is clear that some tools are more disclosure-proof and less rich in information than others. We could draw a line with perfect disclosure protection at one end and complete and full disclosure at the other, with the different tools located somewhere on this line. This book presents evidence that for different data, and in different countries, different choices will be made on this line. The book also shows that these same tools could be ranked in terms of how well they can solve the trade-off dilemma. This situation raises two key questions: Are there tools that are superior to other instruments in that they provide a higher level of confidentiality for the same amount of information richness? Or is it possible to find tools that provide a higher informational content for the same amount of confidentiality?

While we hope this book has gone some of the way toward furthering understanding of the issues involved, more research is needed. Though the research in the book examines information loss in general, it does not address more complex disclosure analysis—for example, disclosure-proofing the results of a set of program benefit simulations when each component of the derivation is slightly adjusted. In addition, different utility metrics, or loss metrics, could be used to quantify disclosure risk and data quality loss. New access modalities, such as simulated access sites, could usefully be explored. Other research, such as research into the area of business perceptions, could be extended and potentially codified by statistical agencies.

The bottom line is that confidentiality is not an arcane topic of little policy interest. Governments and taxpayers pay billions of dollars to statistical agencies to provide decision makers with high quality data. Although there have been no documented cases of disclosing a respondent's identity in the nearly 40 years in which the U.S. Census Bureau, in particular, has released anonymous microdata files, researchers specializing in this area continue to pursue more and more sophisticated anonymization techniques. In addition, disclosure review boards are making increasingly conservative decisions about data release. Users should pay attention to the challenges faced by statistical agencies and work constructively with the agencies to find workable solutions to the core trade-off dilemma. Agencies and users should work together to promote legislative, regulatory, and dissemination policies and practices that facilitate timely and cost-effective access to data for statistical research and policy analysis but do not permit full and open access by all of the public for any use. If confidentiality issues are not fully addressed in constructive and proactive ways, users face the very real risk of losing access to high quality data.

References

Federal Committee on Statistical Methodology (1978) *Report on Statistical Disclosure and Disclosure-Avoidance Techniques (Statistical Policy Working Paper #2)*, Washington, D.C.: U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

—– (1994) *Report on Statistical Disclosure Limitation Methodology (Statistical Policy Working Paper #22)*, Washington, D.C.: U.S. Office of Management and Budget, Statistical Policy Office.