Popul Res Policy Rev (2008) 27:639–665
DOI 10.1007/s11113-008-9095-4

Providing Spatial Data for Secondary Analysis: Issues and Current Practices Relating to Confidentiality

Myron P. Gutmann · Kristine Witkowski · Corey Colyer · JoAnne McFarland O'Rourke · James McNally

Received: 9 July 2005 / Accepted: 18 August 2007 / Published online: 2 July 2008

Abstract Spatially explicit data pose a series of opportunities and challenges for all the actors involved in providing data for long-term preservation and secondary analysis—the data producer, the data archive, and the data user. We report on opportunities and challenges for each of the three players, and then turn to a summary of current thinking about how best to prepare, archive, disseminate, and make use of social science data that have spatially explicit identification. The core issue that runs through the paper is the risk of the disclosure of the identity of respondents. If we know where they live, where they work, or where they own property, it is possible to find out who they are. Those involved in collecting, archiving, and using data need to be aware of the risks of disclosure and become familiar with best practices to avoid disclosures that will be harmful to respondents.

Keywords Archives · Confidentiality · Data · Disclosure · Location

This paper is about the challenges involved in producing, archiving, and sharing social science data that have spatially explicit information embedded within them, all while avoiding the risk of disclosing private information about the individuals who have consented to share information about themselves, in the case of survey research, or who are part of the universe of individuals included in an administrative record system or database. It takes as its starting point the perspective of the data archivist, but it tries to maintain a clear understanding of the competing interests of the data producer, the data user, the survey respondent, and the manager of the data

M. P. Gutmann (✉) · K. Witkowski · J. M. O'Rourke · J. McNally
Inter-University Consortium for Political and Social Research, Institute for Social Research,
University of Michigan, PO Box 1248, Ann Arbor, MI 48106-1248, USA
e-mail: gutmann@umich.edu

C. Colyer
Department of Sociology and Anthropology, West Virginia University, Morgantown, WV, USA

repository, not to mention whatever organization has provided the resources required to collect, clean, document, and disseminate the data. Like others who are concerned about protecting the confidentiality of survey respondents, we are acutely aware that the wealth of information publicly available today increases the risk that someone will breach the promise of confidentiality that is made when most social science data are collected. Spatially explicit data, because they are by definition linked to a specific location that might be someone's home or another easily identifiable place, have the potential to aggravate that risk. Our goal here is to describe many of the issues, identify some of the ways that confidentiality can be protected, and then draw conclusions about current best practices.

Our work arises out of ICPSR's commitment to disclosure analysis and reduction, and an ongoing program of work designed to ensure that the data we archive and disseminate is as safe as possible from disclosure risk (O'Rourke et al. 2006; O'Rourke 2003; Dunn and Austin 1998). This work has led to our involvement in a group of projects at the University of Michigan's Institute for Social Research, designed to help learn more about the risk of disclosure in social survey data, the attitudes that survey respondents have about the risk of confidentiality breach, and current practices within the research community designed to protect survey respondents, and then to ensure that the information we gather is disseminated widely to the research community.[1]

Before we flesh out our discussion, we begin with a simple definition of disclosure as it is associated with social science research data.

For our purposes, we define *Disclosure* as:

> the breach of confidentiality that allows an outsider, whom we call the "intruder," to discern the identity of an individual or business respondent or to associate the individual or business's identity with characteristics reported in a survey or administrative database under a pledge of confidentiality.[2]

Going further, we can define two elements, *attribute disclosure* and *identity disclosure* (Duncan and Lambert 1989; Lambert 1993). *Attribute disclosure* takes as its basic premise that an individual is a respondent in a survey or a subject in an administrative database, and that the intruder knows that the individual is represented in the database. In this case the intruder knows the identity of the respondent but wishes to know specific responses or attributes of that person as recorded in the

[1] Myron Gutmann is principal investigator of NICHD Program Project P01-HD045753, "Human Subject Protection and Disclosure Risk Analysis." The projects are "Informed Consent and Perceptions of Risk and Harm in Survey Participation," directed by Eleanor Singer, Fred Conrad, and Robert Groves, "Estimation of Disclosure Risk and Statistical Methods for Disclosure Limitation," directed by T.E. Raghunathan, Roderick Little, and Richard Valliant, "Disclosure Control: Best Practices & Tools for the Social Sciences," directed by JoAnne M. O'Rourke, and "Resources for the Secure Dissemination of Human Subjects Data," directed by James McNally and Myron Gutmann. The project has a web site at http://www.icpsr.umich.edu/HSP/. We have contributed to related work published as VanWey et al. (2005) and National Research Council (2007), both of which benefited from the fundamental work reported in this paper.

[2] This statement is the authors' working definition, not intentionally a quotation from another source. We believe that the term "intruder" has its origins in the work of Duncan and Lambert (1989). Within the Disclosure literature, data users who attempt to identify anonymized records are called "Intruders" (Elliot 2001) or "Statistical Spies" (Roberts 1986).

database. The intruder attempts to figure out which set of characteristics in the database belongs to the known subject, so that she or he can learn that person's characteristics or attributes. One classic example is that a parent knows that his or her child participated in a school-based survey, but wishes to know the child's response to certain questions, for example about sexual activity or experience with drugs. *Identity disclosure* takes as its basic premise that the intruder does not know that any given individual is a respondent in a survey, but wishes to learn the identity of survey respondents in order to know something about them, to make contact with them, or to harm them or the survey sponsor in some way. Here an example is a marketing firm with a consumer database that it wishes to enrich by identifying and linking information from a large national survey. It would then use its enriched data base to communicate with or sell to those individuals. Another, more pernicious, example would be for the intruder to attempt to identify individuals in a survey merely for the purpose of making their responses known to the general public.

Still more serious, identity disclosure from survey or administrative data might be used by private or public groups to target or harm individuals, population subgroups, or business enterprises. While there are relatively few cases of confidentiality breach by individuals, researchers have found all too many examples of this last form of disclosure risk, whereby groups are identified and harmed using data from official statistics, if not from academic survey research activities (Seltzer and Anderson 2001, 2005, 2007; Anderson and Seltzer 2007). Most analyses of disclosure risk focus on the possibility that an individual may be identified and harmed based on analysis of individual micro-data cases that are publicly released, but this perspective also emphasizes the use of meso- or macro-level data, publicly released or not. In this case, the intruder uses the attributes of a small area (a census tract, for example) to identify the fact that there are individuals in that area who have certain characteristics (an ethnicity, for example), thereby making it worthwhile to target them for repression or other harm. The salient recent example is the use of small area data from the U.S. Census of Population for 2000 to identify areas with large proportions of Arab-Americans after the events of September 11, 2001 (Clemetson 2004; El-Badry and Swanson 2007).

Later we will outline many of the best-known and most widely practiced methods for limiting disclosure risk. There is a growing literature on this topic, in part because both researchers and the statistical agencies of the U.S. government are deeply concerned about the tension between disclosure of public data and protection of confidentiality.[3] The most important element as we begin this discussion is to understand that virtually all widely-used disclosure limitation practices reduce the amount of detail as well as the quality of the information available to the data user. Many of those information reduction activities—for example removing the name of the respondent—have little impact on the analytic value of the data, while others—

[3] The Federal Committee on Statistical Methodology of the U.S. Office of Management and Budget sponsors the Confidentiality and Data Access Committee, which coordinates disclosure management issues for the U.S. government. Its informative web site is at http://www.fcsm.gov/cdac/. This committee is responsible for the Checklist on Disclosure Potential of Proposed Data Releases (http://www.fcsm.gov/committees/cdac/checklist_799.doc), which is widely used by agencies and researchers as a starting point for disclosure review and disclosure limitation.

for example reducing the number of locations or occupational categories preserved in the data in order to eliminate ones that might lead to identification—may reduce the data's analytic value. These procedures all assume that most data are being used for research, and that they can be restricted to use for research and not for other harmful purposes, because restricting those other potential uses requires an entirely different approach, rooted in government policy and public (rather than research) ethics.

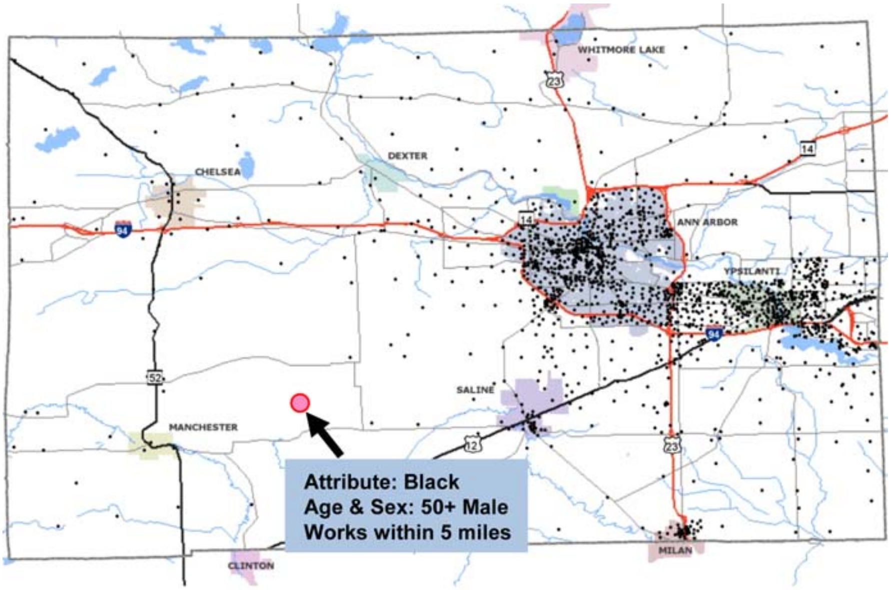
The question that we need to focus on here, then, is how we can maximize analytic value while minimizing the risk of disclosure, and to do so for the special case of spatial information linked to social science data. This is a topic that has not been explored in depth until recently, although there have been two important publications in recent years (VanWey et al. 2005; National Research Council 2007). This article takes a different perspective than either of those, based around the issues faced by data archivists and data users. To make the issue yet more difficult, we need to consider the extent to which it is possible to preserve the concept of public use data (those that are publicly available with little or no restriction on use) while maintaining their research utility. In doing so, we need to think about all the parties to the research enterprise, including data producers, data users, and the respondents whose information we value so highly.

The Data Archivist's Nightmare

The widespread interest in gathering, using, and analyzing spatially explicit information along with social survey data has produced a new nightmare for those of us in the business of preserving and disseminating data. The nightmare stems from a mix of our traditional assumptions about the way that we should do our job and the rapidly changing world of data. We ordinarily assume that with the careful exclusion of direct identifiers (name, address, phone number, social security number), as well as the most common sources of indirect identification (rare occupations or rare places of birth, or rare combinations of these or other variables, for example) from any data that we make public, respondent privacy and confidentiality will be preserved. New technology and the use of spatially explicit data by researchers call these strategies into question.

Our nightmare involves a data producer who deposits data at ICPSR or some other public data repository, and whose data after review are free of all obvious direct and indirect identifiers. As is the case for other data archives and data producers—such as the Census Bureau—who make microdata available, our practice is to provide these data to our community of data users with very limited restrictions. In the case of ICPSR, most such data sets are available for download after the data user agrees electronically to ICPSR's Responsible Use Statement.[4]

[4] ICPSR's Responsible Use Statement is located at http://www.icpsr.umich.edu/org/policies/respuse.html . The data user agrees to these provisions relating to confidentiality: "(1) To use these datasets solely for statistical analysis and reporting of aggregated information, and not for investigation of specific individuals or organizations, except when identification is authorized in writing by ICPSR; (2) To make no use of the identity of any person or establishment discovered inadvertently, and to advise ICPSR of any such discovery; (3) To produce no links among ICPSR datasets or among ICPSR data and other datasets that could identify individuals or organizations."



Attribute: Black
Age & Sex: 50+ Male
Works within 5 miles

Fig. 1 Hypothetical map of Washtenaw County, Michigan, revealing the characteristics of a single respondent, with the location of that respondent's home

This is all fine, but the nightmare risk comes in the case that the researcher who collected the data reveals other information about respondents, especially when those data are revealed spatially. Even quite simple information linked to spatially explicit identifiers can be dangerous, for example a map that is part of a presentation, publication, or web site. A potential intruder may be able to derive an address (or address range) from a published or presented map, use the address to narrow the potential group of households that might include the respondent, and then use other information in the map or presentation to link to the public access data.

Figure 1 represents hypothetical mapped data for Washtenaw County, Michigan. Initially, it might just be a map of points representing the residences of respondents to a hypothetical survey.[5] This particular map—which we might imagine appearing in a researcher's presentation, publication, or web site—tells us that at one location we know that the respondent is a male African American, aged over 50, who works within five miles of his residence. Those are straightforward characteristics, and on their own they are unlikely to carry a substantial disclosure risk in most parts of the U.S. What would be frightening for everyone involved would be if someone matched this location to a street address map and pinpointed this respondent to a single house or more likely (given the scale) to a residential block. That makes the risk of disclosure much greater, because even these generic characteristics might lead to identification if the universe from which we choose is so small. If the intruder goes one step further and finds this case in the public data file, the risk

[5] It is important to note that addresses are never available for U.S. census data, and could never be mapped in this way. This is not necessarily the case for other data sources, however, where members of the data production team might have access to addresses and believe that mapping them would be useful.

becomes much greater. Not only are there many more attributes in the public data file that might make it possible to identify the survey respondent, but there may be responses to other, more sensitive questions in the public data file. The data producer may be willing to take a chance that an intruder can identify an individual and know that he is male, black, 50 and over, and works within five miles of his residence, but questions about income or personal behavior may reveal something that the respondent did not want others to know. This could seriously compromise not only the study in question, but the whole process by which social scientists convince their research subjects to reveal information (Singer et al. 1993).

This nightmare exists because researchers have changed their approach to managing, collecting, and analyzing data, and they have done so for the good reason that they can answer important questions by doing so (Rindfuss 2002). Researchers wish to make use of information such as location in their analysis; this by design has an increased risk of disclosure. Most of the data collected are not new: data producers have always had to know the identity of their respondents, including name, address, telephone number, or school, and possibly their social security or driver's license number, the location of their employment, or the places where they own agricultural or other property. Until the change that we document here, few data producers had the means to release spatially explicit information (for example through maps), and extremely few potential intruders had the technical expertise to use what information was released to identify survey respondents. Moreover, until fairly recently data producers and secondary data users were content to operate within a system that tried to ensure that data subject to long-term preservation and used for secondary analysis had all direct identifiers removed, including locations as well as name, address, telephone, and social security number.

There is a good reason why researchers wish to make use of spatial data: it has the potential to produce answers to important questions. Among the earliest known models for this kind of productive research is John Snow's famous study of the London Cholera epidemic of 1854 (see Fig. 2), where he mapped the prevalence of disease and the proximity of disease to the pumps that London's population used to obtain drinking water (Snow 1855; Tufte 2001).[6] The pump closest to the most cholera cases was the likely source of the disease. The value of the cartographic analysis is in direct proportion to the precision with which the researcher knows the location of the disease outbreak. Yet that precision also has the potential to reveal information about the individuals or households whose attributes are included in the map. The growing use of spatially explicit information derives not only from advances in social science theory, but in technological changes, including the increased availability and usability of Geographical Information Systems, and the increased reliance on hierarchical and multi-level statistical models that often call for information about spatial location in order to insert the individual or household in its appropriate context, for example a block, census tract, or city (Armstrong 2002; Goss 1995; Armstrong and Ruggles 2005; Boulos et al. 2006; Brownstein et al. 2006).

[6] For an excellent summary of Snow's work and its impact, see this web site maintained by the UCLA School of Public Health: http://www.ph.ucla.edu/epi/snow.html.

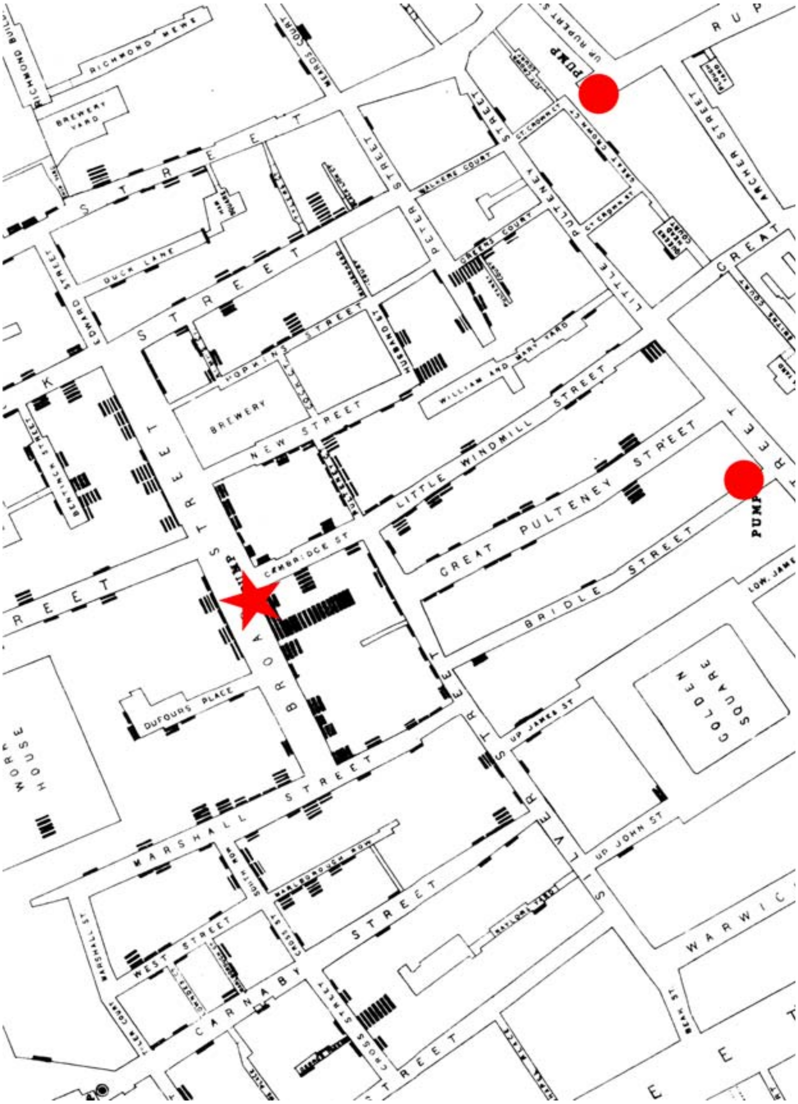


Fig. 2 John Snow's map of the 1854 cholera outbreak in London. Note that the map only solved the problem because of unique locations (Snow 1855)

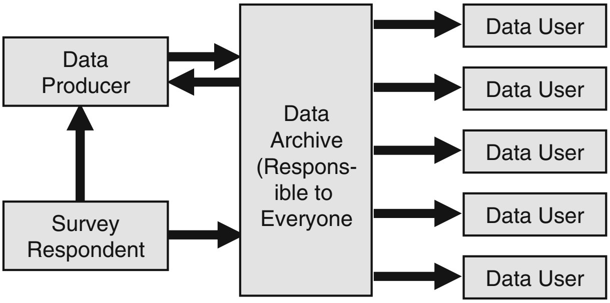


Fig. 3 A simple model of actors and actions in the public data world

Much is new in spatial social science research, despite its antecedents. Since the 1990s (if not earlier), social scientists have been gathering information as part of their surveys that is spatially explicit and integral to the analytic core of their research. Researchers who study the relationship between population, land use, and environment have asked respondents about the location of their agricultural plots, and then surveyed those plots and recorded their location. They have also linked those locations to remotely-sensed satellite and aerial photo data.[7] These data give their research an unequalled spatial richness that is resonant with that of the classical works of epidemiology, and has similar potential to answer important research questions.

The tension revealed by the data archivist's nightmare is a conflict between two concepts. On the one hand, disclosure limitation procedures are designed to eliminate those attributes or combinations of attributes that make any respondent unique. We work to eliminate uniqueness. On the other hand, as in the case of Snow's map, the locations in spatial data are unique by definition, and their value comes from their uniqueness. In this case, we work to preserve uniqueness. This conflict can also partly be revealed as reflecting the interests of the parties whose interests are at play in terms of spatial data and disclosure.

Competing Needs in the World of Data

The data world is populated by individuals and groups who have competing needs and goals. In the simplest way (Fig. 3), we can represent three classes of individuals with a professional interest in data (data producers, data archivists, and data users), as well as the survey respondent, who is a crucial participant but not professionally involved. Each of these parties has its own view of the world of data dissemination, with its own needs and sources of anxiety. Data producers are involved in the process of data production because they see a need to answer important research questions; that need is primary. At the same time, they are increasingly pressured—both by research funders and by professional values—to release their data for use by

[7] Some of this work is well summarized in two National Academy Publications (Liverman et al. 1998; Entwisle and Stern 2005), and especially in Moran and Brondizio (1998) and Entwisle et al. (1998).

others. They also sometimes face with ambivalence the conflict between a goal of maintaining complete control over their data and an alternative approach that hands the data over to the data archive's specialists in the preservation and dissemination of data. Data users are not a totally separate group, because they overlap in many cases with data producers. Data users, on the other hand, need access to data that can serve their research questions without being subject either to restraints on access to data or to limits on the analytic utility of the data. Survey respondents have been brought into the research process because information about them has value for important societal questions. Researchers promise them confidentiality and preservation from unnecessary risk of harm in order to secure their consent to participate (Singer 1993, 1978; Singer et al. 1992, 1995). These promises must be kept, ensuring the long-term protection from harm of survey respondents. The data archivist attempts to serve all the needs just listed, and more. We are responsible for helping the data producer preserve and share his or her data, for helping the data user get relatively easy access to data that will answer the research questions that they have identified, and for constantly being on the watch to preserve the confidentiality promises given to the survey respondent.

Meeting all these needs by and for all the parties to the research process has become a constantly moving target and an ever more complicated process, especially over the past decade. At the very heart of the problem is the increasing quantity of information about individuals that is available to the public, or available with relatively little difficulty and at relatively low cost (Sweeney 2001). On-line telephone directories are the simplest of these publicly available and searchable databases. Others are more restricted, but they have much more information, such as proprietary (but still widely available) marketing and credit databases (such as Experian—http://www.experian.com), and other more public sources, such as the Aristotle database (http://www.aristotle.com) that is largely built from voter registration data, are available to anyone who is willing to pay a fee. One only has to spend a short time searching the web to know that there is a great deal of information readily available.[8] The danger of public release of information is made more complicated by the rapid increase in computing power available to all, which means that efforts at data matching and data mining that would have been unattainable to most intruders a few years ago are no longer outside the realm of possibility. Spatial data add to this problem by making it easier for the intruder to refine the search based on geographical constraints.

At the same time that it has become easier for the intruder, the public—and public officials—have become more sensitive to the risk of disclosure (Robbin 2001).[9] This sensitivity has increased the resistance of potential respondents to participation in surveys. Throughout the U.S. the public discussion of identity theft

[8] Searching any of the commonly used online directories (whitepages.com, netscape.com, etc.) for a name yields a link to a service that will do a background check on that individual, as well as perform other kinds of searches, all for a moderate fee. For an example, see http://www.publicdata.com/.

[9] On March 1, 2005, the U.S. Census Bureau named Gerald W. Gates its first "Privacy Officer." The Bureau's concern about public perceptions and their impact on doing its job is evident in the web page entitled "Are you in a survey," which is part of the Bureau's Data Protection and Privacy Policy web page: http://www.census.gov/privacy.

has grown in volume and frequency, leading to increasing caution about the sharing of information.[10] The general discourse about these issues has been heightened by a number of legislative initiatives.[11] There have been a steady stream of efforts since the 1970s to protect privacy (the Privacy Act of 1974), keep student and other public records from being disclosed (the Federal Family Educational Rights and Privacy Act of 1974), and culminating recently with the extensive requirements of the Health Insurance Portability and Accountability Act of 1996 (PL 104-191), otherwise known as HIPAA, and the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA).[12] While it is important to emphasize that the most damaging public releases of information have come from intrusions into confidential administrative databases held by universities, government agencies, and private businesses, and not from statistical or research data, the discussion has led to heightened public anxiety about all forms of data collection and data maintenance.[13]

This is where the intruder, the last party to our public data world, enters the picture (Fig. 4). By adding the intruder to our picture, we see that in many ways the data files become the central element in our conceptualization, with a variety of actors surrounding them. Each of these actors—the human subject, the various research interests, and the intruder—have their own goals with regard to the data and their own relationship with the other parties. And each is constrained differently by legal and regulatory constraints, no more so than the actors who are committed to conducting and facilitating research.

The identity of intruders into social science data remains a mystery to us, which is probably a good thing because it means that there have not yet been enough intrusions into these research data to warrant serious study. It is likely—but not at all certain—that such intruders are different from the criminals who are known to exploit commercial databases in order to commit identity theft. We presume that one category of intruders are those who know someone is involved in a study, perhaps an angry spouse or a curious parent who feels that they have sufficient reason to tackle data in order to discover important information. We presume that another category of intruder is a "hacker"—someone who has an interest in understanding technology and a grudge against the system—who attacks a data source merely because he can and because he enjoys disruption. There is limited evidence about the existence of both of these categories of intruders, but it is limited

[10] There are a large number of discussions of identity theft in the news media and on public agency web sites, as well as special-purpose organizations that purport to work to thwart identity theft, but the quantity of serious academic and legal discussion of the problem thus far is quite limited.

[11] The University of California Libraries have a good web site on Privacy legislation and issues at http://libraries.universityofcalifornia.edu/privacy/legis.html#federal.

[12] The Privacy Act of 1974 (PL 93-579-5 U.S.C. 552a); the Federal Family Educational Rights and Privacy Act of 1974 (FERPA), dated July 17, 1976 (20 U.S.C. Section 1232g-34 CFR Part 99); and the Health Insurance Portability and Accountability Act of 1996 (PL 104-191); Confidential Information Protection and Statistical Efficiency Act of 2002 (PL 107-347-44 USC 3501).

[13] In February and March, 2005, the major story has been the unintentional sale of data by the data brokers ChoicePoint and Lexis-Nexis to criminals who apparently used the information for identity theft. See, for example the multitude of stories in the New York Times, or Bruce Schneier's weblog: http://www.schneier.com/blog/archives/2005/02/choicepoint.html.

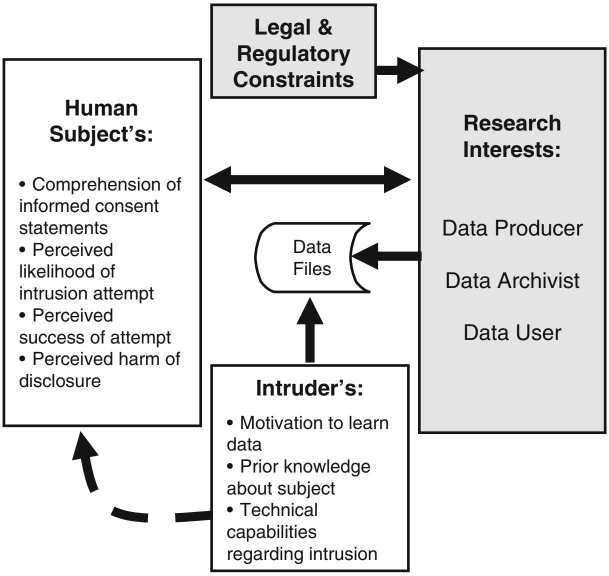

Legal & Regulatory Constraints
Human Subject's:
• Comprehension of informed consent statements
• Perceived likelihood of intrusion attempt
• Perceived success of attempt
• Perceived harm of disclosure
Research Interests:
Data Producer
Data Archivist
Data User
Data Files
Intruder's:
• Motivation to learn data
• Prior knowledge about subject
• Technical capabilities regarding intrusion

Fig. 4 A more complex model of actors and actions in the public data world

for good reason: data producers and repositories rarely disclose any evidence they might have of intrusions or intruders, or even whether such intrusions have taken place. Our limited knowledge about intrusions means that we can only speculate about intruders, their practices, and their motives.

Disclosure Limitation Basics

The body of literature on disclosure limitation methodology dates to the 1970s and has become quite large.[14] The field emerged in the late 1960s as computer scientists discovered new efficient means to link administrative records (Fellegi 1972). In 1976, the Federal Committee on Statistical Methodology (FCSM) formed a subcommittee to investigate ways to release statistically useful information without subjecting respondents to confidentiality risks. In 1978, the committee issued its first report, "Statistical Policy Working Paper 2: Report on Statistical Disclosure and Disclosure Avoidance Techniques" (Subcommittee on Disclosure-Avoidance Techniques 1978). Working Paper 2 set forth a framework for defining statistical disclosure and summarized the techniques available for limiting disclosure risk in data products. Although updated during the past quarter-century (Subcommittee on Disclosure Limitation Methodology 1994), these techniques continue to be used as the foundation of disclosure limitation practice today.

[14] Duncan (2001) provides a brief and effective overview, with a bibliography. For a fuller (but obviously not complete) bibliography on the topic, see O'Rourke and Gutmann (2005), on-line at http://www.icpsr.umich.edu/HSP/citations/index.html.

The principles laid out in Working paper 2 and in important subsequent work by many authors include: (1) Releasing samples from larger populations; (2) Reducing variability of categorical data where the attributes facilitate identification of specific sub-populations in the data; (3) Suppressing data that render unique cases visible; and (4) Stochastically perturbing individual values in a manner that retains the original covariate structure. In technical terms disclosure limitation involves manipulating or masking a data matrix. Following important work by Cox (1980), Dalenius and Reiss (1982), Roberts (1986), Dalenius (1986), and Duncan and Lambert (1986), Duncan and Pearson (1991) provide a clear and accessible description of the methods available for the reduction of disclosure risk. More recent monographic treatments by Willenborg and de Waal (1996, 2001) and Doyle et al. (2001) update the framework to emphasize disclosure risk assessment and administrative alternatives to statistical solutions.

In this section we provide an overview of disclosure limitation and attempt to establish some ground-rules for understanding how disclosure risk can be minimized and what such minimized disclosure risk means for subsequent analysis. Here, we are concentrating on disclosure limitation in situations where spatial data are not explicitly in play.

Our goal in limiting disclosure risk is to minimize the visibility of unique respondents. The simplest and most dangerous case of a unique occurs when an individual in a survey has a combination of characteristics that would make him or her unique in the general population, thus making the respondent easily identifiable. Consider an individual who has a unique occupation within a large geographic area, such as the single member of the U.S. House of Representatives within some states. Somewhat less dangerous occurrences arise when a respondent has unique characteristics within the survey (but not necessarily in the general public), and when there are attributes of the survey design that make it possible to identify that individual. An example of this kind of unique would be a study with a highly clustered design (for example, schools), where it was possible to find out about the sampling frame. It would not be a surprise, for example, that students, parents, teachers, and staff, not to mention other members of the community, had learned that a given school was a primary sampling unit for a school-based survey. Highly clustered sample designs pose serious disclosure risks, both because they provide so many opportunities for there to be unique cases, and because so many individuals are likely to know of the study's existence. Most forms of explicit spatial data also pose serious disclosure risks.

We limit disclosure risk first by eliminating all obvious identifiers, whether they be direct or indirect. A direct identifier is an attribute like a name, social security number, telephone number, and so on. Indirect identifiers are attributes which are not necessarily exclusive to the individual in question, but can be used in combination with other measures to unambiguously identify. For instance, in a demographic dataset, the date of birth and date of death taken together with county of residence can work as indirect identification mechanisms. After addressing obvious identifiers, we focus on other cases that have unique combinations of key attributes. Data from these cases can be modified or transformed in a number of ways to confuse potential intruders.

$$X = \begin{matrix} n_1 p_i & \cdots & \cdots & n_1 p_4 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ n_4 p_1 & \cdots & \cdots & n_4 p_4 \end{matrix}$$

Fig. 5 Disclosure limitation techniques: matrix masking of non-spatial data. Consider the dataset X to be an *n*-by-*p* matrix. Each of the *n*-rows provides information on *i* subjects for *p* attributes

We explain this in more detail later, but the key is to change the data matrix (by adding, subtracting, or modifying cases) in order to prevent the intruder from knowing with certainty if re-identification occurred.[15] What may be most important in thinking about disclosure limitation is that the data producer consider these issues as close to the beginning of the project as possible. A dissemination strategy is not sufficient if it only consists of a statement to the sponsors of the research that says that the investigator will share the data. Rather, it is essential that the investigators plan for dissemination from the very start of their research, and possibly even earlier. Many things can go wrong if not planned for early. Early planning helps everyone involved in the process make good decisions about the ultimate dissemination of a data collection and helps prevent inadvertent mistakes that can complicate dissemination. The publication of maps with individual cases shown is one of them. Appropriate informed consent procedures and discussions with local Institutional Review Boards for human subjects are another. Moreover, it is not good practice to announce to the public or the research community the location of research sites that can be thought of as primary sampling units, despite how grateful we are that a school, hospital, location, or other site has agreed to participate. That information alone is enough to compromise a data sharing plan for a project, and limit our ability to share the data in the future.

Conceptually, in order to understand how to think about disclosure risk and disclosure limitation, we should see the data in a set of survey responses as an *n*-by-*p* matrix (Fig. 5). Each of the *n*-rows provides information on *i* subjects for *p* attributes. The rows represent records for individual respondents, while the columns represent the values provided by the respondent to each question. Key attributes are those that when combined can identify a respondent, even when identification requires information coming from other, external sources. Disclosure risk is reduced by eliminating unique sets of key attributes, and by adding noise to data elements (Kim 1986; Kim and Winkler 2003). We accomplish this by changing either records, or attributes, or both.

Records can be changed by making public only a sample of the original cases, so intruders cannot be sure that someone they know in the sample is actually represented by the record identified (Fig. 6a). An example is a parent of a child enumerated in a school-based study. The parent knows that the child is in the overall sample because he signed a release, but he may not be able to be certain that he has

[15] For example, the public use files from the National Longitudinal Study of Adolescent Health (Add health) that are distributed by the Sociometrics Corporation (www.socio.com) only contain a randomly chosen one-in-two sample of the original cases.

A. Releasing only a sample of the data (subtracting rows from X)

$$\text{X sampled} = \begin{matrix} del & del & del & del \\ n_2 p_1 & \ddots & n_2 p_3 & \vdots \\ \vdots & & \ddots & \vdots \\ n_4 p_1 & \dots & \dots & n_4 p_4 \end{matrix}$$

B. Dropping sensitive variables from the file (subtracting columns from X)

$$\text{X -sp} = \begin{matrix} n_2 p_1 & \dots & \dots & del \\ \vdots & \ddots & & del \\ \vdots & & \ddots & del \\ n_4 p_1 & \dots & \dots & del \end{matrix}$$

C. Altering sensitive cells through recoding, collapsing, top or bottom coding, or adding stochastic error (perturbing X)

$$\text{X perturbed} = \begin{matrix} n_2 p_1 & \dots & \dots & del \\ \vdots & \ddots & n_2 p_3' & \vdots \\ \vdots & n_3 p_2' & \ddots & \vdots \\ n_4 p_1 & \dots & \dots & n_4 p_4 \end{matrix}$$

D. Swapping or relocating attributes from one record to another

$$\text{X swapped} = \begin{matrix} del & del & del & del \\ n_4 p_1 & n_4 p_2 & n_4 p_3 & n_4 p_4 \\ \vdots & & \ddots & \vdots \\ n_2 p_1 & n_2 p_2 & n_2 p_3 & n_2 p_4 \end{matrix}$$

Fig. 6 Disclosure limitation techniques: matrix masking of non-spatial data. Four specific approaches

identified the correct respondent in the sub-sample because some cases have been eliminated. Alternatively, we can add synthetically generated records to multiply the number of cases that share key attributes, eliminating uniques. In this case we create a situation where every unique in the original sample was represented by two or more cases, making it impossible to know which one of them is actually the case the intruder has in mind. A third alternative for record changing is to displace records spatially (Fig. 6d). In this transformation, we swap records from one place to another, so that information from an individual with a certain set of key attributes is exchanged with the information from another individual, located in a different geographical area, but who matches the same attributes. (Dalenius and Reiss 1982; Fienberg and McIntyre 2004; Takemura 2002; Zayatz 2003). In this case, the intruder believes that he has identified an individual he knows, but attributes that are not shared between the two cases are not necessarily the same. They may have the same age, sex, race, occupation, and marital status, but their response on income, attitudes, or other behaviors may be different. The intruder believes that he has identified a case, but he is mistaken in the information that he draws from the process.

Transforming attributes is a different process, although it shares characteristics with record-changing procedures, specifically with record swapping. In general, attribute transformations require us to change the values of one or more of the attributes of an individual case in the study in order to reduce the likelihood of their being identified by values in the data. The simplest way to do this, although it can represent a substantial reduction in the analytic value of the data, is to eliminate sensitive variables completely (Fig. 6b). Tail-censoring (of which top-coding is a special case) is another category of attribute transformation, where rare values in the tail of a distribution are grouped together so that none of them exceeds a certain range, for example recoding incomes over a certain amount so that all those individuals have as their income the cut-off (say by recoding all incomes over $100,000 to "$100,000 or more"). Tail-censoring or top-coding is a special case of all forms of perturbation, including recoding, collapsing, bottom coding, and adding stochastic error (Fig. 6c). Some of the other widely-used approaches are micro-aggregation, where individual values are clustered into small aggregates of size k, and then replaced with the average value from the aggregate (Domingo-Ferrer and Mateo-Sanz 2002; Domingo-Ferrer et al. 2002; Sande 2002), and the post-randomization method (PRAM), where scores from categorical variables are changed to a different score prescribed by a probability mechanism.

The best efforts at disclosure limitation make use of a wide range of expertise, drawn not only from the producers of the data, but also data users, statisticians, and experts on data dissemination and disclosure (O'Rourke 2003). When done effectively, it retains statistically useful variables and does not change the statistical properties of the data. In reality, changes to the statistical properties may be inevitable, and at the very least these changes may not always be fully understood in advance (Little 1993; Domingo-Ferrer and Torra 2001; Armstrong et al. 1999). A further limitation is that many disclosure limitation procedures have a disproportionate impact on the analysis of small sub-populations that may be of high research interest, for example racial minorities, pregnant women, and people with rare health problems. Because these individuals are infrequent in the population, they may require more data transformations to prevent disclosure, thereby further complicating statistical analysis.

As a practical matter, researchers and data archivists have learned that they cannot always limit disclosure risk enough so that data may be made available to the general public, even under data use limitations like those imposed by ICPSR. Instead, the most effective practitioners of data sharing for sensitive data have developed schemes whereby they categorize data into as many as four or five categories, depending on the sensitivity of the data and the risk of disclosure after masking procedures have been completed.[16] Low risk data can be shared widely. In the case of sensitive data or data with higher disclosure risk, restricted use

[16] See, for example, the restricted data protections for the Los Angeles Family and Neighborhood Survey (L.A. FANS) at http://www.lasurvey.rand.org/data/restricted; the National Longitudinal Adolescent Health Survey (Add Health) at http://www.cpc.unc.edu/projects/addhealth/data; and the Health and Retirement Survey (HRS) at http://hrsonline.isr.umich.edu/rda/reslis2.php.

dissemination may be more desirable. An example would be when the data user (and his or her employer) is required to implement data protection technology and sign a contract ensuring the promise of non-disclosure. In the most sensitive cases, these data may only be made available in a data enclave, like those at major data centers like the Carolina Population Center, ICPSR, the Institute for Social Research at the University of Michigan, the National Center for Health Statistics' Research Data Center, and the network of Census Research Data Centers maintained by the U.S. Bureau of the Census.[17] These more restrictive solutions—especially those involving enclave protection—are an important solution but a last resort for all concerned because they severely restrict access to data for many potential data users who have difficulty relocating to the site of the enclave.

Recent research has also led to insights into an alternative to restrictive solutions, in the form of synthetic data (Little 1993; Rubin 1993; Raghunathan et al. 2003; Abowd and Lane 2004; Reiter 2002, 2005). Conceptually, synthetic data can be created by modeling the attributes of a given research data set, and then using those modeled attributes to create a synthetic population with the same statistical characteristics. Because the cases in the data file are all created by the "synthesizer," they can be publicly disseminated without any disclosure risk at all: all of them are made up in the modeling process, and none of them are actual human research subjects. Rubin (1993) argues that datasets generated through multiple imputation procedures can actually provide more useful information to an analyst than "real" data coarsened through more traditional disclosure techniques. While it is not yet clear to what extent synthetic data will be useful for all research questions, they constitute an important starting point for thinking about ways of making data public while still preserving the confidentiality and privacy of survey respondents.

Disclosure Limitation and Spatial Data

Spatial data add a wealth of opportunities for social science research, while simultaneously adding disclosure risk because of their very precision. In this discussion we do not attempt to differentiate between the growing variety of forms of spatial data that are available or may be available. Some of these data are as simple as the exact address or latitude and longitude of a house, school, medical facility, or place of work, obtained during survey data collection. Other forms of spatial data come from various kinds of remotely-sensed sources, including aerial photographs, satellite imagery, and the output from devices that help ascertain and record an individual's location, such as a cell phone or the toll-paying tags now used in many localities. From our point of view the source of information is less relevant than the fact that one or more precise locations can be associated with data of use to social science researchers.

[17] For the NCHS Research Data Center, see http://www.cdc.gov/nchs/r&d/rdc.htm. For the Census Research Data Center program and centers (there are currently 9, with a 10th opening soon) see http://www.ces.census.gov/index.php/ces/researchprogram.

We begin our discussion of ways to deal with spatially explicit data by emphasizing that most mechanisms for limiting disclosure in data where locational precision is not required involve collapsing geographical categories. The widely-used Public-Use Microdata Samples (PUMS) produced by the U.S. Bureau of the census for the 2000 enumeration of the United States are a case in point (U.S. Bureau of the Census 2003).[18] In these data, for example, no area with a population smaller than 100,000 persons is identified in the 5% sample, and none with a population smaller than 400,000 is identified in the 1% sample.[19] By collapsing geography to these categories, and then by invoking a series of other disclosure-reduction measures, the Census Bureau ensures that there are no unique individuals in any of the geographic areas it reports.[20] The PUMS, of course, is drawn from what is already a very large sample, consisting of approximately 15.8% of U.S. residents, those who completed the long form of the census questionnaire. Smaller samples can require more substantial collapsing of geographical units, and might report geography at a larger scale, for example the state or region, or in some instances, without any geographical precision at all, besides reporting that all of the respondents live within the United States.

Aggregating geography has the desirable effect of easily reducing disclosure risk, but it does so at great cost, because it limits the precision with which we know where people live, or work, or transact other activities in their lives. We see in Figs. 7 and 8 the dramatic impact of reducing spatial precision, here by collapsing our imagined Washtenaw county population (divided into Black and a non-Black respondents) to the scale of census tracts. Figure 8 is interesting, but it can’t give us the detail shown in Fig. 7, and if we were to collapse even more thoroughly—to the city, county, or state—we would have a very different kind of geographical information, with much less precision. Spatial precision, as we learned from John Snow’s work, can have important analytic value. The key to defining best practices for public use of spatially explicit data will be to find the most effective compromise between masking data by aggregating geography so that data can be made public, and restricting data so that the maximum analytic value can be preserved. We need to keep as much information as possible while protecting respondents from breach of confidentiality, at the same time not losing sight of the special characteristics of spatial information: it is useful because it is precise.

When we turn to strategies for diminishing disclosure risk for spatially explicit data, we find ourselves with three major alternatives, each designed to reduce the level of risk while maintaining as much analytic utility as possible (Armstrong et al. 1999). We can summarize these approaches as ones that adjust the coordinates, ones that aggregate points, and ones that attach contextual data prior to dissemination.

[18] The areas of 100,000 or more in the PUMS are called Public Use Microdata Areas (PUMAs), and the areas of 400,000 are called super-PUMAs. The full range of disclosure avoidance techniques used by the Census Bureau are described in the PUMS Technical Documentation (U.S. Bureau of the Census 2003), and in Hawala (2003). For the debate that led to these classifications, see Robbin (2001).

[19] The 100,000 rule is embedded in the FCSM Checklist (Interagency Confidentiality and Data Access Group, Federal Committee on Statistical Methodology 1999).

[20] Zayatz (2002); Zayatz et al. (1996). For other U.S. agencies, see Jabine (1993). For the more general question, see Duncan et al. (1993).

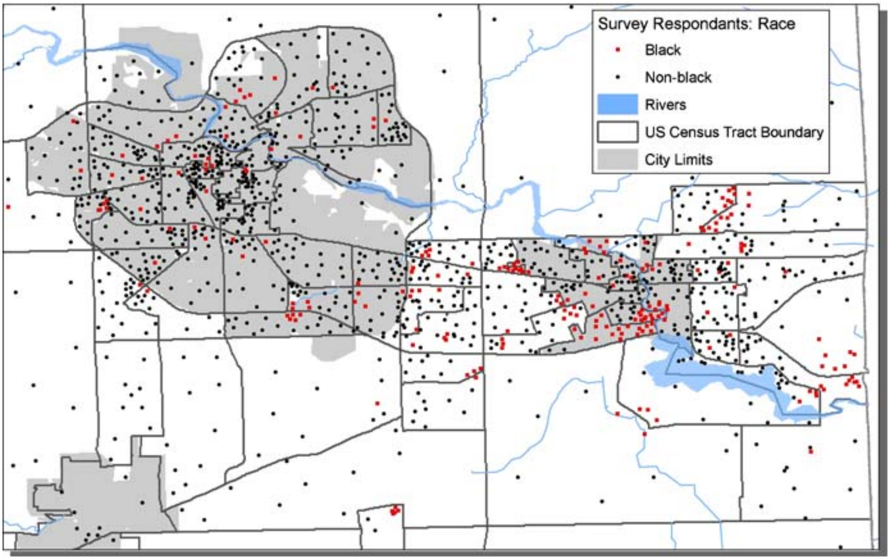


Fig. 7 Highly detailed data about hypothetical respondents to a survey

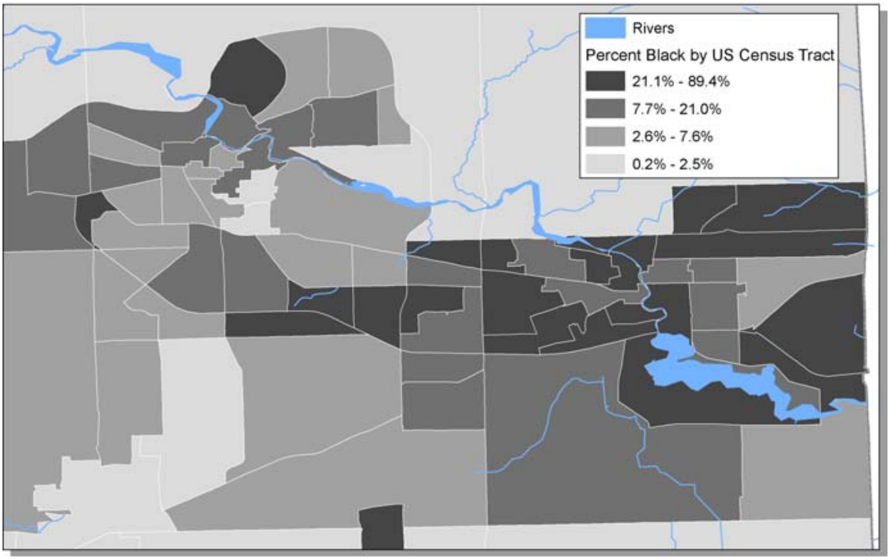


Fig. 8 Data about hypothetical respondents to survey, aggregated to census tract

All of these can be successful at masking the identity of respondents, and all do so by reducing some of the potential data use or analysis that can be done. The problem posed by spatial data makes these solutions both necessary and obvious. When we attach specific geographic coordinates to social survey data, we provide those with access to the data with an absolute means to recognize the location at which the activity (living, working) takes place. Only by changing those coordinates in some way can we limit disclosure, but by changing or masking those coordinates we reduce their precision. The key is to make those changes while still retaining as much of the core information represented by the coordinates as possible.

Adjusting Coordinates

The simplest way to minimize the risk of disclosure in the case of data with exact spatial coordinates is to displace those coordinates in some way. The risk in doing so, of course, is that displacement may distort the spatial information that makes the data valuable in the first place. Among the techniques used to adjust coordinates, the first and most straightforward is to displace points, either by shifting them with a fixed or random increment. We can, for example, add or subtract a fixed or random amount to both the *x*-axis and the *y*-axis. Other alternatives that fall in the same family of procedures are to change the scale that represents the relationship between points, to rotate them, or to use some combination of displacement, scale change, and rotation.[21] Adjusting coordinates in a systematic way may preserve internal spatial relationships, so that a study designed to measure the distances family members travel to assist each other may maintain those distances, but the more systematic the adjustment, the more likely that an intruder can reverse the process and decode the original locations. Combining adjustment procedures makes it easier to mask location, but at the same time makes analysis more difficult. It may also be necessary to individualize data sets with adjusted coordinates, each designed to optimize analytic capacity for a given research application while minimizing the risk of disclosure. Yet creating a multitude of individualized data sets based on a single original study increases the risk of confusion among data users and makes it possible—at least in theory—for an intruder to combine customized but public data sets and reverse-engineer the masking used in them.

Aggregation

An alternative to adjusting coordinates is to aggregate points. At its simplest, aggregation is exactly what researchers have always done to mask location and minimize the possibility of disclosure. Instead of representing the exact location, we collect and summarize values for an administrative unit, such as a census tract, county, city, state, or even country. There are also more sophisticated ways of doing this, which are fundamentally similar to other kinds of micro-aggregation. If we believe that the distance between relatively close-by points is not essential to the analysis, we can take a fixed or varying number of nearby points and give them the same location, either that of one of the points, or some synthesized location (for example the centroid of the polygon implied by the original points). This has the potential to maintain certain information for analysis at a scale less than that of an

[21] It is usually not possible to perform these adjustments automatically, because it is often necessary to ensure that the adjusted coordinates do not conflict with physical realities, for example by putting a house in a lake, or a road that mounts straight up a hill or mountain.

arbitrary administrative unit, while still masking exact locations. It is less useful if the exact spatial relationship of one location to the others is essential to the analysis.

Attaching Contextual Variables

Sometimes the only reason that researchers require spatial information is in order to be able to attach contextual variables. We want to know, for example, the characteristics of the population (or some other set of attributes) of the administrative unit in which the individual lives, rather than the exact location of that residence or the distance from that residence to place of work or health care. By providing the data user with the exact location of the respondent's home we give that data user the flexibility to merge individual-level data with whatever contextual data are most appropriate. At the same time, this strategy raises the risk of disclosure. The alternative is for the data producer to attach a set of contextual variables to the individual-level data, while removing the precise location from the data set.[22] This has the advantage of preserving confidentiality while giving the data user a range of contextual variables with which to enrich his or her analysis. On the other hand, it reduces the data user's flexibility, because she or he is dependent on the contextual variable choices of the producer. And unless done carefully, it does not necessarily eliminate disclosure risk. Depending on sample design and the contextual variables attached, the contextual variables themselves may constitute a unique geographic definition, and allow an intruder to identify the location of the contextual unit, for example a census tract or county. It is easy to imagine that a combination of population size broken down by sex and race might uniquely identify an area the size of a census tract, at least within a known state or census region. If it were possible to match a respondent to an area that small, the task of identification based on other attributes would be much easier. The same is true if the respondent were identified by area code and telephone exchange, especially given the newly published demographic data available at that resolution, as well as at the resolution of the nine-digit zip code.[23]

This discussion of spatial data and the means that might be used to mask location produces a few key findings. The first is that it is possible to preserve spatial relationships within data while reducing disclosure risk, but the task is not an easy one. Either we reduce the analytic value of the data or we increase the risk of disclosure. If the attributes of interest can be preserved by aggregating locations, either through ordinary aggregation or micro-aggregation, or by adding contextual variables and eliminating the precise locations, the task is much simpler—again at the expense of losing spatial precision. Whatever approach we take, Armstrong and his colleagues have shown that different approaches allow different kinds of spatial analysis, implying the creation of more than one public-use data set to represent a single original study. That may be satisfactory, but there is an unmeasured risk

[22] This is the practice of the National Survey of Family Growth (NSFG), for example.

[23] These data are available from Geolytics, Inc., a producer of high quality and widely used demographic products. The risk posed by these products with demographic attributes and differing geographies, when coupled with other attributes, is well-described in Steel and Sperling (2001).

created by putting multiple data sets in the public domain, because it may be possible to combine them and reverse-engineer the masking in order to discover the original locations. Spatial data are also prime candidates for restricted-use data agreements and enclave protections, with the same warning that all such restrictions sharply diminish the availability of research data to many potential users and thereby limit the return on investment. Synthetic data are an enticing prospect for spatial data, but relatively little work has been done in this area, and the value of synthesizing data with specific spatial attributes is still unknown.

Challenges for the Actors

Early in this paper we identified a group of actors whose roles are important in understanding disclosure risk. Before we turn to best practices and recommendations it is worthwhile returning to them and their concerns, and to highlight the challenges each faces, especially in the situation where spatially explicit data have been created and have the potential to be shared with other researchers.

Data producers need to be applauded for their determination and creativity in coming up with ways to collect and analyze data with so much more valuable information than has been previously available. Spatial data are a wonderful tool that enhances research possibilities for everyone. At the same time, data producers need to understand from the very beginning of the research process that there are risks associated with sharing their data, and they need to work with determination to overcome those risks without giving up and without falling back on needlessly restrictive practices that make data sharing difficult for the rest of the community. They need to embrace and support innovative work in the area of data transformation and synthetic data creation if those new ideas will lead to wider use of their data. Similarly, they need to form partnerships with the data archiving and preservation community to ensure that they are aware of the most up-to-date techniques for ensuring that their data are preserved and made available for others. Their other close partners are survey respondents, who need to be protected in a rapidly changing world, but who also are required to take seriously the informed consent that they give.

Secondary data users need to be aware of the risks of disclosure and the obligations that data producers and data archivists have taken on in the area of respondent protection. They need to understand that spatially explicit data and data with other kinds of highly clustered sample designs pose special disclosure risks that require special protection and treatment. Their challenge is to find ways to work with data that have in some cases been transformed in ways that make them less useful for secondary analysis than they were to the primary analyst, or that have restrictions on use that make them less than optimally accessible. Data users have an obligation to engage with all the other actors in a constructive way to make sure that as much research as possible is done, with as little risk of harm to respondents as possible.

We in the archiving community have the challenging role of being in the middle of all the action, with the responsibility to protect the interests of all the other parties: producers, users, and respondents. Our challenge is to be pro-active in seeking out the best ways to preserve data and make them available, to communicate these best practices to data producers, and then to document and explain the resulting decisions to data users. Spatial data make that task more difficult, because our world is increasingly changing from one where our only obligation was to distribute totally safe and anonymous data to all who wanted it, and where we knew how to ensure that safety. We now are part of a world where opportunities to identify respondents grows daily, and where spatial information makes that ever easier. Studies that were safe to distribute just two or three years ago may no longer be safe to distribute without a refreshed disclosure review, or without restrictions that make it difficult for data users to complete their research.

Best Practices and Recommendations

In addressing this moving target, part of our long-term research project involves a survey of data producers funded by federal agencies, in which we ask them about their practices for protecting the confidentiality of respondents. We will use the results of that survey to continue to develop a list of best practices. We summarize our discussion by emphasizing four main points.

(1) It is essential that the data producer think about dissemination planning as early in the research process as possible, and consider questions of disclosure risk at every stage in the research design. This is especially important in the case of spatial data because the inadvertent release of certain kinds of spatial information can diminish the ability to make other forms of data publicly available later.

For example, at the stage at which the investigator is applying for funds, she or he should inform the potential funder that spatially explicit data are involved, and that they cannot easily be released as anonymous public data, but may require some form of restricted data release. These forms of data release may lead to additional costs, which need to be born by one party or another, either funder, data producer, data user, or data archive. In addition, it is crucial to reveal to Human Subjects review boards that location data will be collected or merged with social survey data, and that data dissemination is planned. It is equally important to include this notice in the informed consent process. At the same time, it is reasonable to say that the investigators will not publicly release identities or locations associated with survey respondents.

At later stages in the process, it is equally important that researchers keep disclosure risk in mind, especially for spatial data. We have shown the dangers that follow from the publication of maps that identify individual respondents along with their responses to individual survey questions. Any disseminated information that reveals primary sampling units in highly clustered designs makes public release sensitive, whether that information reveals the primary sampling unit directly (by giving its name) or indirectly (for example by publishing characteristics of the PSU that would allow it to be identified).

(2) There exists a large body of well-known practices that mask identities in data and limit disclosure risk. These practices—briefly introduced and documented in this paper—constitute the starting point for all assessments of disclosure risk and

implementations of disclosure limitation strategies. Those who work on data sharing are well advised to use those procedures as an explicit checklist, starting with the work of the Subcommittee on Disclosure Avoidance Techniques (1978, 1994), continuing with the work of Duncan and Pearson (1991), and including the recent recommendations from Armstrong et al. (1999); Doyle et al. (2001); and Willenborg and de Waal (2001). The Checklist on Disclosure Potential of Proposed Data Releases (Interagency Confidentiality and Data Access Group 1999) is a starting point for the verification of these processes, which our research on best practices is designed to complement and enhance. Those procedures that allow public data dissemination often limit the analytic utility of the data, however, especially for spatially explicit data.

The simplest example of the loss of analytic utility comes from removing spatial information from the data file, especially if that removal takes away information that was crucial to the study design (sample weights) or to the analysis that the original research team conducted. Eliminating spatial variables used by the initial investigators in their research may protect respondents, but it does not permit secondary analysis or replication, which is the goal of data sharing. Effective data sharing designs need to protect the needs of respondents while understanding the needs of secondary data users, and not focus exclusively on the needs and desires of data producers.

(3) One size does not fit all types of data and all types of users. Different data users sometimes require different types of data, with different attributes from a disclosure risk perspective. As Armstrong et al. (1999) show, it is possible to share spatial data in more ways if the data producer is willing to transform the data differently for different users, for example by performing microaggregation for some while rotating the locations of points for others. The notion of customizable distribution of data is an enticing prospect for the future, especially if we find a way to ensure that multiple versions of the same underlying data do not lead to potentially risky situations.

(4) The data producer and the data archivist need to be sensitive to the wide variety of ways that data can be disseminated effectively, not all of which involve masked public-use data sets. Other options, including restricted-use contracts, the deposit of data in protective enclaves, and the enticing but not yet fully developed concept of synthetic data sets are all options. Where precise information on location is crucial to successful research, these options may be the only ones available.

Following those best practices will go a long way towards preventing our data disclosure nightmare from ever coming to pass. The nightmare does not require that someone has made spatial data public, only that whatever spatial information (maps, for example) is made public be sensitive to ways that it can be used by an intruder to locate respondents and then link them to non-spatial information that is made public. Avoiding the nightmare requires a great deal of advance thought on the part of the data producer about what she or he can publish without risking disclosure of the identities of respondents, and about the fact that new technologies and systemic weaknesses in our data systems make protecting the confidentiality a game with moving targets. It is essential that we be proactive in ensuring as much protection as possible.

Acknowledgments This research has been supported by Grant Number P01 HD045753 from the National Institute of Child Health and Human Development. We are grateful to assistance from our colleagues Elizabeth Moss for work on the bibliography, to Lisa Isgett for the maps, and to Emily Klancher Merchant for assistance with the text. Robert Groves helped with the original thinking embedded in Figs. 3 and 4. We presented an earlier version at Indiana University in November, 2004, and at the Annual Meeting of the Population Association of America, Philadelphia, April, 2005.

References

Abowd, J. M., & Lane, J. (2004). New approaches to confidentiality protection: Synthetic data, remote access and research data centers. In J. Domingo-Ferrer & V. Torra (Eds.), *Privacy in statistical databases* (pp. 282–289). New York: Springer-Verlag.

Anderson, M., & Seltzer, W. (2007). Challenges to the confidentiality of U.S. federal statistics, 1910–1965. *Journal of Official Statistics, 23*(1), 1–34.

Armstrong, M. P. (2002). Geographic information technologies and their potentially erosive effects on personal privacy. *Studies in the Social Sciences, 27*(1), 19–28.

Armstrong, M. P., & Ruggles, A. (2005). Geographic information technologies and personal privacy. *Cartographica, 40*(4), 63–73.

Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine, 18*(5), 497–525.

Boulos, M. N. K., Cai, Q., Padget, J. A., & Rushton, G. (2006). Using software agents to preserve individual health data confidentiality in microscale geographical analyses. *Journal of Biomedical Informatics, 39*(2), 160–170.

Brownstein, J. S., Cassa, C. A., & Mandi, K. D. (2006). No place to hide—reverse identification of patients from published maps. *New England Journal of Medicine, 355*(16), 1741–1742.

Clemetson, L. (2004). Homeland Security given data on Arab-Americans. *The New York Times*, July 30, p. A14.

Cox, L. H. (1980). Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association, 75*(370), 377–385.

Dalenius, T. (1986). Finding a needle in a haystack or identifying anonymous census records. *Journal of Official Statistics, 2*(3), 329–336.

Dalenius, T., & Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference, 6*(1), 73–85.

Domingo-Ferrer, J., & Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering, 14*(1), 189–201.

Domingo-Ferrer, J., Oganian, A., Torres, A., & Mateo-Sanz, J. M. (2002). On the security of microaggregation with individual ranking: Analytical attacks. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems, 10*(5), 477–491.

Domingo-Ferrer, J., & Torra, V. (2001). Disclosure control methods and information loss for microdata. In P. Doyle, J. I. Lane, J. J. M. Theeuwes & L. M. Zayatz (Eds.), *Confidentiality, disclosure, and data access* (pp. 91–110). Amsterdam: North-Holland.

Doyle, P., Lane, J. I., Theeuwes, J. J. M., & Zayatz, L. V. (Eds.). (2001). *Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies*. Amsterdam: North-Holland.

Duncan, G. T. (2001). Confidentiality and statistical disclosure limitation. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 2521–2525). Oxford: Elsevier Science Ltd.

Duncan, G. T., Jabine, T. B., & de Wolf, V. A. (Eds.). (1993). *Private lives and public policies: Confidentiality and accessibility of government statistics*. Washington, DC: National Academy Press.

Duncan, G. T., & Lambert, D. (1986). Disclosure-limited data dissemination (with discussion). *Journal of the American Statistical Association, 81*(393), 10–18.

Duncan, G., & Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business & Economic Statistics, 7*(2), 207–217.

Duncan, G. T., & Pearson, R. W. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science, 6*(3), 219–232.

Dunn, C. S., & Austin, E. W. (1998). Protecting confidentiality in archival data resources. *ICPSR Bulletin, 19*(1), 1–8.

El-Badry, S., & Swanson, D. A. (2007). Providing census tabulations to government security agencies in the United States: The case of Arab Americans. *Government Information Quarterly, 24*(2), 470–487.

Elliot, M. (2001). Disclosure risk assessment: Confidentiality, disclosure, and data access. In P. Doyle, J. I. Lane, J. J. M. Theeuwes & L. M. Zayatz (Eds.), *Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies* (pp. 75–90). Amsterdam: North-Holland.

Entwisle, B., & Stern, P. (Eds.). (2005). *Population, land use, and environment: Research directions*. Washington, DC: National Academy Press.

Entwisle, B., Walsh, S. J., Rindfuss, R. R., & Chamratrithirong, A. (1998). Land-use/land-cover and population dynamics, Nang Rong, Thailand. In D. Liverman, E. F. Moran, R. R. Rindfuss & P. C. Stern (Eds.), *People and pixels: Linking remote sensing and social science* (pp. 121–144). Washington, DC: National Academy Press.

Fellegi, I. P. (1972). On the question of statistical confidentiality. *Journal of the American Statistical Association, 67*(337), 7–18.

Fienberg, S. E., & McIntyre, J. (2004). Data swapping: Variations on a theme by Dalenius and Reiss. In J. Domingo-Ferrer & V. Torra (Eds.), *Privacy in statistical databases* (pp. 14–29). Berlin: Springer-Verlag.

Goss, J. (1995). We know who you are and we know where you live: The instrumental rationality of geodemographic systems. *Economic Geography, 71*(2), 171–198.

Hawala, S. (2003). *Microdata disclosure protection research and experiences at the U.S. Census Bureau*. Paper Presented at the Workshop on Microdata, Stockholm, Sweden. http://www.census.gov/srd/sdc/microdataprotection.pdf. Cited 18 Feb 2007.

Interagency Confidentiality and Data Access Group, Federal Committee on Statistical Methodology. (1999). *Checklist on disclosure potential of proposed data releases*. Washington, DC: Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget. http://www.fcsm.gov/committees/cdac/.

Jabine, T. B. (1993). Statistical disclosure limitation practices of United States statistical agencies. *Journal of Official Statistics, 9*(2), 427–454.

Kim, J. J. (1986). A method for limiting disclosure in microdata based on random noise and transformation. In *American Statistical Association, Proceedings of the Section on Survey Research Methods* (pp. 303–308).

Kim, J. J., & Winkler, W. E. (2003). *Multiplicative noise for masking continuous data*. Census Statistical Research Report Series: RRS2003/01. http://www.census.gov/srd/papers/pdf/rrs2003-01.pdf.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics, 9*(2), 313–331.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics, 9*(2), 407–426.

Liverman, D., Moran, E. F., Rindfuss, R. R., & Stern, P. C. (Eds.). (1998). *People and pixels: Linking remote sensing and social science*. Washington, DC: National Academy Press.

Moran, E. F., & Brondizio, E. (1998). Land-use change after deforestation in Amazonia. In D. Liverman, E. F. Moran, R. R. Rindfuss & P. C. Stern (Eds.), *People and pixels: Linking remote sensing and social science* (pp. 94–120). Washington, DC: National Academy Press.

National Research Council. (2007). In M. P. Gutmann & P. Stern (Eds.), *Putting people on the map: Protecting confidentiality with linked social-spatial data*. Washington, DC: National Academy Press.

O'Rourke, J. M. (2003). Disclosure analysis at ICPSR. *ICPSR Bulletin, 24*(1), 3–9.

O'Rourke, J. M., & Gutmann, M. P. (2005). Citations database—human subjects protection and disclosure risk analysis, project 3: Statistical disclosure control: Best practices and tools for the social sciences. Ann Arbor, MI: Inter-university Consortium for Political and Social Research. http://www.icpsr.umich.edu/HSP/citations/index.html.

O'Rourke, J. M., Roehrig, S., Heeringa, S., Reed, B. G., Birdsall, W. C., Overcashier, M., et al. (2006). Solving problems of disclosure risk while retaining key analytic uses of publicly released microdata. *Journal of Empirical Research on Human Research Ethics, 1*(3), 63–84.

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics, 19*(1), 1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics, 18*(4), 531–543.

Reiter, J. P. (2005). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society Series A, 168*(1), 185–205.

Rindfuss, R. R. (2002). Conflicting demands: Confidentiality promises and data availability. *Newsletter of the International Human Dimensions Programme on Global Environmental Change Update*, 2. http://www.ihdp.uni-bonn.de/html/publications/update/IHDPUpdate02_02.html.

Robbin, A. (2001). The loss of personal privacy and its consequences for social research. *Journal of Government Information, 28*(5), 493–527.

Roberts, H. V. (1986). Disclosure-limited data dissemination: Comment. *Journal of the American Statistical Association, 81*(393), 25–27.

Rubin, D. B. (1993). Satisfying confidentiality constraints through the use of synthetic multiply-imputed microdata. *Journal of Official Statistics, 9*(2), 461–468.

Sande, G. (2002). Exact and approximate methods for data directed microaggregation in one or more dimensions. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems, 10*(5), 459–476.

Seltzer, W., & Anderson, M. (2001). The dark side of numbers: The role of population data systems in human rights abuses. *Social Research, 68*(2), 481–513.

Seltzer, W., & Anderson, M. (2005). On the use of population data systems to target vulnerable population subgroups for human rights abuses. *Coyuntura Social, 32*, 31–44.

Seltzer, W., & Anderson, M. (2007). *Census confidentiality under the Second War Powers Act (1942–1947)*. Paper presented at the Annual Meeting of the Population Association of America, New York. http://www.uwm.edu/~margo/govstat/Seltzer-AndersonPAA2007paper3-12-2007.doc. Cited 15 Jul 2007.

Singer, E. (1978). Informed consent: Consequences for response rate and response quality in social surveys. *American Sociological Review, 43*(2), 144–162.

Singer, E. (1993). Informed consent in surveys: A review of the empirical literature. *Journal of Official Statistics, 9*(2), 361–375.

Singer, E., Hippler, H.-J., & Schwarz, N. (1992). Confidentiality assurances in surveys: Reassurance or threat. *International Journal of Public Opinion Research, 4*(3), 257–268.

Singer, E., Mathiowetz, N. A., & Couper, M. P. (1993). The impact of privacy and confidentiality concerns on survey participation: The case of the 1990 U.S. Census. *Public Opinion Quarterly, 57*(4), 465–482.

Singer, E., Von Thurn, D. R., & Miller, E. R. (1995). Confidentiality assurances and survey response: A review of the experimental literature. *Public Opinion Quarterly, 59*(1), 66–77.

Snow, J. (1855). *On the mode of communication of cholera* (2nd ed.). London: John Churchill.

Steel, P., & Sperling, J. (2001). *The impact of multiple geographies and geographic detail on disclosure risk: Interactions between census tract and ZIP code tabulation geography*. Paper presented at the annual meeting of the American Statistical Association, Survey Research Methods Section.

Subcommittee on Disclosure-Avoidance Techniques, Federal Committee on Statistical Methodology. (1978). *Statistical policy working paper 2: Report on statistical disclosure and disclosure avoidance techniques*. Washington, DC: U.S. Department of Commerce.

Subcommittee on Disclosure Limitation Methodology, Federal Committee on Statistical Methodology. (1994). *Statistical policy working paper 22: Report on statistical disclosure limitation methodology*. Washington, DC: Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

Sweeney, L. (2001). Information explosion: Confidentiality, disclosure, and data access. In P. Doyle, J. I. Lane, J. J. M. Theeuwes & L. M. Zayatz (Eds.), *Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies* (pp. 43–74). Amsterdam: North-Holland.

Takemura, A. (2002). Local recoding and record swapping by maximum weight matching for disclosure control of microdata sets. *Journal of Official Statistics, 18*(2), 275–289.

Tufte, E. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire: Graphics Press.

U.S. Bureau of the Census. (2003). *Census 2000, Public use microdata sample, (PUMS), technical documentation*. Washington, DC: U.S. Census Bureau.

VanWey, L., Rindfuss, R., Gutmann, M. P., Entwisle, B., & Balk, D. (2005). Confidentiality and spatially explicit data: Concerns and challenges. *Proceedings of the National Academy of Sciences of the United States of America, 102*(43), 15337–15342.

Willenborg, L., & de Waal, T. (1996). *Statistical disclosure control in practice*. New York: Springer-Verlag.

Willenborg, L., & de Waal, T. (2001). *Elements of statistical disclosure control*. New York: Springer-Verlag.

Zayatz, L. (2002). SDC in the 2000 U.S. decennial census. In J. Domingo-Ferrer (Ed.), *Inference control in statistical databases* (pp. 183–202). Berlin: Springer-Verlag.

Zayatz, L. (2003). *Disclosure limitation for census 2000 tabular data*. Paper presented at the Joint European Commission for Europe and EUROSTAT Work Session on Statistical Data Confidentiality. http://www.unece.org/stats/documents/2003/04/confidentiality/wp.15.e.pdf.

Zayatz, L., Moore, R., & Evans, B. T. (1996). *New directions in disclosure limitation at the Census Bureau*. Census Statistical Research Report Series: LVZ96/01.