Privacy Protection in Social Science Research: Possibilities and Impossibilities

Jeremy J. Albright, *Methods Consultants of Ann Arbor, LLC*

ABSTRACT The ubiquity of data in the twenty-first century provides unprecedented opportunities for social science research, but it also creates troubling possibilities for privacy violations. The emerging field of statistical disclosure control (SDC) studies how data collectors and analysts can find an optimal solution to balancing privacy protection and data utility. This article introduces SDC to readers in the applied political science research community and outlines its implications for analyzing individual-level data. The vocabulary of SDC is introduced and is followed with a discussion emphasizing just how easy it is to break almost any release of supposedly "anonymized" data. The article then describes how SDC measures almost always destroy the ability of researchers to accurately analyze complex survey data. These results are in conflict with increasing trends toward greater transparency in the social sciences. A discussion of the future of SDC concludes the article.

> *"You have zero privacy anyway. Get over it."*
>
> —Scott McNealy, CEO, Sun Microsystems
> January 25, 1999

In the fall of 2006, the popular online DVD rental service Netflix announced a competition to help improve its ability to make movie recommendations to subscribers. Most Netflix users record their movie preferences using a five-star rating system, and the company decided to leverage its extensive ratings database to develop better predictive models of subscriber tastes. Toward this end, a subset of the data was released to the public along with an announcement of a $1 million prize to be awarded to the person or team who developed the most accurate recommendation algorithm. Recognizing that many subscribers would not want their preferences publicly known, Netflix promised that the data file had been stripped of all potentially identifying information before release. In early 2008, however, two researchers from the University of Texas announced that they had breached the anonymization of the Netflix Prize dataset and, using external information collected from member profiles on the Internet Movie Database website, were able to learn about subscribers' political and religious preferences (Narayanan and Shmatikov 2008).

The story of the Netflix Prize is a prominent example of the paradox of twenty-first century research. Scholars are awash in vast amounts of data, which—along with the enormous storage capacity and processing power of modern computers—make it possible to study social phenomena in unprecedented ways. At the same time, the amount of information that researchers, governments, and private companies collect about individuals raises troubling questions about privacy (National Research Council 2007). Indeed, as the Netflix Prize database demonstrates, it may be possible to recover subject identities even when overt identifiers have been altered or removed entirely. Variables that, on their own, seem benign may be used in combination to identify a respondent, while one dataset may be merged with another to gain sufficient auxiliary information to locate unique subjects (Bethlehem, Keller, and Pannekoek 1990; Jaro 1989). The result is that guaranteeing the protection of privacy in any data collection project is far more ambiguous than many people realize.

In the face of increasing concerns with privacy protection, the field of statistical disclosure control (SDC) developed to quantify

Jeremy J. Albright *is founder and president of Methods Consultants of Ann Arbor, a statistics consulting firm. Among its services, the company offers advising on and audits of data privacy measures. He can be reached at jeremy@methodsconsultants.com.*

doi:10.1017/S1049096511001405

https://doi.org/10.1017/S1049096511001405 Published online by Cambridge University Press

both the risk of subject reidentification and the effects of privacy protection measures on data utility (Willenborg and de Waal 2001). There are two reasons why political scientists ought to pay attention to this literature. First, the findings have direct implications for how social science researchers share their results, as evolving norms of transparency and replication lead more and more researchers to post their data and syntax files online (Freese 2007; King 1995; McDermott 2010). SDC provides guidelines for determining levels of disclosure risk present in a data file, but unfortunately most publications related to SDC appear in journals from the fields of statistics and computer science. This article describes the SDC literature and its application in the context of political science research. Second, privacy promises to be a fundamental political issue in the twenty-first century. The classic tension between regulation and innovation will play out over issues of data access. Recent controversies about Facebook privacy settings and Google's inadvertent collection of private data represent potential harbingers of future controversies. Debates about the political ramifications of privacy issues require an understanding of the technical issues involved with SDC. It is hoped that this article contributes to such an understanding.

The article develops as follows. The first section introduces the field of SDC by defining key terms, describing how researchers quantify risk, identifying options to minimize risk, and outlining how these decisions affect the usefulness of a data file. A subsequent section describes the implications of SDC for political science research, viz. the problems it introduces for variance estimation in complex surveys. A final section outlines where the field of SDC is headed.

THE FIELD OF STATISTICAL DISCLOSURE CONTROL

The Netflix Prize deanonymization is not the only recent case illustrating threats to individual privacy in the modern, data-rich era. In the late 1990s, a privacy researcher was able to merge information from voter registration files with a database she purchased from the Group Insurance Commission of Massachusetts to uniquely identify health records for Massachusetts state governor William Weld (Sweeney 1997, 2002). In 2006, AOL released web search data from 658,000 (supposedly deidentified) users for the benefit of any interested researcher. It was quickly determined, however, that many search histories could easily lead to reidentification and reveal sensitive or embarrassing information (Hansell 2006). That same year, geographers used minimally detailed maps published in newspapers to reengineer the exact locations of houses in which bodies were recovered following Hurricane Katrina (Curtis, Mills, and Leitner 2006). The findings, the authors noted, have implications for any epidemiological map displaying disease outbreaks. In 2009, two researchers from Carnegie Mellon University used data from the Social Security Administration's Death Master File to develop a statistical model that predicts with great accuracy—on the basis of knowing only date and place of birth—the social security number of anyone born in the United States (Acquisti and Gross 2009). Those results demonstrate that privacy risks are pervasive in a data-rich environment, including risks to individuals who have not contributed any information to a particular statistical database. Indeed, the social security study exemplifies an impossibility result from the SDC literature—discussed later—showing that absolute privacy protection can never be guaranteed.

Sweeney (2000) has suggested that it is possible to uniquely identify about 87% of the US population on the basis of gender, birth date, and zip code alone. This information is easily merged with publicly available data sources, such as voter registration files, to determine the proper names of subjects. Given that an abundance of data raises concerns for privacy, data providers take careful steps to minimize the risk of disclosure through masking potentially sensitive variables before public release. Yet, it is well known that many of the most commonly used masking techniques diminish the analytic usefulness of the information (Bialik, 2010). Therefore, finding the optimal trade-off has led to the emerging field of statistical disclosure control (SDC).

The Vocabulary of SDC

In 1977, statistician Tore Dalenius suggested that privacy protection in databases should be defined according to the amount of new information one can learn about an individual after seeing a data file. It was not for another decade, however, before the field of SDC developed an identity of its own. In the mid-1980s, a handful of researchers began proposing early frameworks for thinking about risks to the privacy of study subjects (Duncan and Lambert 1986). Before then, statisticians and computer scientists acknowledged the presence of disclosure risk in microdata files (data files with information collected on individuals), but little accumulation of knowledge existed in the field. For example, computer scientists had been working for decades to find ways to merge cases from different databases in which the matching variables were imperfectly recorded (Fellegi and Sunter 1969; Jaro 1989). However, these linkage algorithms were presented as helpful tools for improving information systems rather than as a potential means for breaching privacy. At the same time, data collection agencies such as the US Census Bureau recognized the potential for inadvertent disclosure through the release of both microdata files and data summary tables. Nonetheless, there was little guidance on how to alter data files in a manner that would retain the usefulness of the information (Willenborg and de Waal 2001).

The work of Duncan and Lambert provided both a vocabulary and framework for thinking about disclosure risk (Duncan and Lambert 1986, 1989; Lambert 1993). They defined an *intruder* (sometimes also called an *adversary*) as a person or organization seeking to identify a record or multiple records in a data file. The intruder has a certain amount of prior information about the records that, on viewing the released data, is updated to form a posterior assessment of the probability that a case in the file can be matched to a real person. A wide range of possible scenarios describe both intruders' prior knowledge and their motivation for attempting an identification. An intruder may know for certain that a case is in the released file, may suspect with some probability that the person is present, or may not know of any cases in the file but simply wants to identify anyone to discredit the data collecting agency (Paaß, 1988). The amount of risk in the Duncan and Lambert framework therefore depends on assumptions about an intruder's knowledge combined with the amount of information in the released data.

The variables that an intruder will attempt to use to identify a subject are known as *keys*, which are sometimes also called *quasi-identifiers* because they do not explicitly reveal an individual's name but can be used to deduce or infer subject identities. It is never sufficient to consider each key variable on its own, as key values are most revelatory when they are combined to produce sparsely populated cells in an *m*-dimensional table (for *m* keys). For example, variables measuring gender, occupation, income, and state of

https://doi.org/10.1017/S1049096511001405 Published online by Cambridge University Press

residence may seem benign in isolation. However, if a case in a data file happened to correspond to, say, a wealthy female public servant from Alaska, the number of possible matches would decrease dramatically. Thus, every variable that can be used as a key must be identified and all of the possible value combinations must be examined.

Before looking at how researchers have attempted to formalize disclosure risk, however, the next section describes steps that statisticians have developed to make privacy breaches more difficult. These steps will be introduced first because, as the subsequent section will show, they still fail to provide absolute security and may entail unacceptably high levels of degradation to data quality.

Masking Sensitive Data

One of the earliest approaches to masking sensitive data, and one which is still used today by organizations such as the United States Census Bureau, is known as *data swapping* (Reiss 1984; Zayatz 2006). Data swapping consists of interchanging values on key variables between similar units, thereby introducing greater uncertainty into an intruder's attempts to claim a definitive identification. One common approach to swapping is to switch cases across different geographic units, although swapping can take place with reference to nonspatial data as well.

The formulas used to carry out data swaps are designed to ensure that relationships between the swapped variables are retained. That is, the covariances among all the swapped variables will be approximately the same both before and after exchanging values. This makes it possible to test multivariate hypotheses when only swapped variables are included in the model *or* when only nonswapped variables are included. Estimation becomes problematic, however, when relationships between the swapped and nonswapped variables are explored, as the original relationships are not necessarily retained.

A second technology comes from the development of algorithms for *microaggregation*, a process in which continuous or categorical variables are optimally collapsed into broader categories and then assigned the average score from those assigned groupings. The goal of microaggregation is to coarsen the data so that the cells in m-dimensional tables constructed from the key variables are more highly populated. This should be done, however, in a manner that minimizes the loss of variance in the masked variables (Defays and Anwar 1998; Domingo-Ferrer and Mateo-Sanz 2002). The microaggregation algorithm determines the optimal cut-points for the different groupings, and innovations in microaggregation allow the process to be extended to account for multivariate relationships. Despite the optimization of threshold decisions, however, the resulting masked data always contain less information than the original, which in turn attenuates the researcher's ability to make firm statements about relationships between variables.

Another classic approach to data masking is to supplement observations with random noise (Doyle et al. 2001; Duncan and Mukherjee 2000). Adding a stochastic component to observations using any commercial statistical package is easy, and the resulting error in the masked variables again obfuscates a record's true identity. Nonetheless, there are two major limitations to the random perturbation approach. First, the amount of noise required to ensure protection may be substantial, which in turn can dramatically alter relationships between perturbed and nonperturbed variables. Second, and more important, the security gains may be illusory for basic applications of additive noise, making even more sophisticated perturbation methods necessary, which in turn increases the burden on the data provider (Kargupta and Datta 2003).

Another approach to data protection is simply suppressing (that is, recode as missing) observations and variables. The μ-Argus SDC software system contains algorithms for optimally suppressing only certain high-risk cells in a data file (Hundepool et al. 1998; Willenborg and de Waal 2001). Cell suppression, as opposed to deleting variables or cases in their entirety, allows retention of a maximal amount of information in a data file while guaranteeing that unique key combinations are removed. The cost, of course, is that some information is still lost, even if the amount of missing data has been minimized. Similar methods and trade-offs exist for top- and bottom-coding variables (e.g., recoding all incomes above $100,000 to be $100,000).

Finally, a recent innovative alternative has been the release of entirely synthetic data files simulated to look like the original. Because this approach is still in early development, it is examined in the final section of this article.

Defining and Measuring Risk

Attempts to measure risk have gone hand-in-hand with attempts to provide an acceptable definition of privacy. Thus, different *types* of privacy have been proposed along with algorithms to provide a data file whose contents are consistent with the authors' particular privacy definition. Only recently has a relatively robust form of privacy protection been proposed, but it requires a different way of thinking about doing research.

An early and intuitive privacy definition was k-anonymity (Sweeney 2002). A microdata file achieves k-anonymity when combinations of the key variables occur at least k times in the records. The higher the value for k, the more difficult it is for an intruder to claim with certainty to have made a correct linkage. A data agency can set a threshold for an acceptable level of k-anonymity and, if a file does not yet meet the threshold, the data curating agency can mask or suppress observations until k-anonymity has been reached. Although k-anonymity captures the problem of uniqueness among observations, k-anonymity is an incomplete guarantee of privacy protection.

Machanavajjhala et al. show two ways to break k-anonymity (Machanavajjhala et al. 2006). In the first, a group of individuals within one subgroup sharing identical values on the key variables may also be homogenous on a sensitive attribute. If k-anonymity is set at eight, and all eight individuals in a group have the same illness, then it is possible to deduce that an individual has an illness even if his or her exact record is not known. In a second attack, the intruder uses information about the distribution of characteristics in the population to locate an individual's record. For example, more than one illness may be listed within a grouping of k individuals. However, additional information, such as race or ethnicity, can rule out some of the values if an illness is rare in the population for similar individuals.

Machanavajjhala et al. thus introduce an alternative model of privacy termed ℓ-diversity, which stipulates that each grouping of variables (created, for example, with microaggregation) has ℓ "well-represented" values of the sensitive variables (Machanavajjhala et al. 2006). The meaning of "well-represented" varies depending on different instantiations of the ℓ-diversity principle. The authors show that meeting ℓ-diversity provides protection against both

https://doi.org/10.1017/S1049096511001405 Published online by Cambridge University Press

types of attacks they described in the context of k-anonymity. Yet, shortly after ℓ-diversity was introduced, it too was breached by multiple researchers. Li, Li, and Venkatasubramanian show that ℓ-diversity is problematic for skewed distributions on sensitive variables, such as would obtain if an illness (the sensitive attribute) were present in less than 5% of the population (Li, Li, and Venkatasubramanian 2007). Li, Li, and Venkatasubramanian suggest t-closeness as an alternative criterion, which specifies that the distribution of sensitive values within a subgrouping should reflect its distribution in the overall table. In a separate paper, Xiao and Tao show that neither k-anonymity nor ℓ-diversity provide much protection to dynamic databases whose cases are added and deleted at different times (Xiao and Tao 2007). They propose a concept called m-invariance as another alternative.

Recognizing that a vexatious break-propose-break cycle was developing in the field of SDC, several researchers at Microsoft developed a more formalized definition of privacy. Their definition clarifies what it means to compromise anonymized data without falling back on the kinds of ad hoc formulations offered with each new SDC break (Dwork et al. 2006). Dwork (2006) illustrates the difficulty of this task by proving an impossibility result. She shows that absolute privacy—in the terms suggested by Dalenius in 1977—could never be obtained due to the omnipresence of external sources of information. Because complete disclosure control is impossible, Dwork argues in favor of tying measures of privacy protection to the amount of risk that exists to an individual for participating in a database. Although such a definition gives up on the hope for absolute privacy protection, it nonetheless addresses data collectors' concern that a privacy breach will discourage potential subjects from participating in future studies. Reticence vis-à-vis study participation can be ideally reduced if subjects are convinced that taking part will not pose much of a privacy threat beyond what exists without participating.

Dwork's privacy principle is termed *differential privacy*, and its key innovation is to define privacy not in terms of the database—which essentially all prior work in the field had done—but to define it in terms of the types of queries made against the database. Privacy definitions tied to the database itself will always be breachable given some additional knowledge, whereas defining privacy in terms of the function queried against the dataset eliminates the need to worry about auxiliary information. A query made against the data can achieve a predetermined level of differential privacy by introducing an element of random noise to the result of the query (as opposed to adding noise to each observation, as previously described). The amount of noise depends, in turn, on the sensitivity of the function, which is defined as the maximal amount of change possible in the function's value after removing any single case from the database.

Two major drawbacks to differential privacy have limited its widespread use. First, except for very simple count queries, differential privacy requires a large amount of noise addition, so that inferences become highly unreliable (Muralidhar and Sarathy 2010). Second, differential privacy requires a different understanding of handling data. For example, social scientists are accustomed to downloading a data file and having the raw data available to manipulate and query. Differential privacy only works if the microdata are not accessible and if software for differential privacy protecting queries exists that is both widely accessible and user friendly. Such software is a long way off, and its widespread use requires a reorientation of how researchers interact with data. Thus, most suppliers of social science data will continue to use masking techniques to meet the more rudimentary privacy definitions.

Measures of Data Utility

If achieving k-anonymity or a similar measure of privacy were possible and robust to privacy breaches, the amount of data alteration required to reach the specified level of security may be substantial. Consequently, data users will be very interested in knowing how much useful information remains in a masked data file. Several measures of data utility thus have been proposed to help inform end users.

Domingo-Ferrer and Torra (2001a,b) summarize the measures for assessing the utility of both continuous and categorical variables as being based on the idea that a masked file should look like the original as much as possible. As one option, they propose estimating mean square or mean absolute errors between the original and masked values. For non-continuous variables, Domingo-Ferrer and Torra suggest categorical analogs to distance-based measures, such as the maximum number of categories that fall between a masked ordinal variable and its original value. They also suggest a measure based on Shannon's entropy (Gomatam and Karr 2003; Willenborg and de Waal 2001), a concept developed in communication theory to measure the amount of distortion in a signal that is sent over a noisy channel.

Willenborg and de Waal (2001) provide a thorough explanation of the intuition behind adapting Shannon's entropy to quantify information loss. The intruder's goal is to reconstruct an original dataset, but the intruder only observes a noisy signal about the content of the original file by examining the released (masked) data. From the released data and some knowledge of the masking procedure, the intruder considers the probabilities of different possible original datasets that could have been generalized to produce the one dataset that is publicly available. The entropy of the distribution with the highest probability, H (Old|New), is then taken to be the measure of information loss.

Gomatam and Karr also provide a comprehensive list of information loss metrics for discrete variables in addition to entropy, including an approach based on computing distances between the joint distributions of the pre- and post-swapped data files (Gomatam and Karr 2003). These options do not exhaust the available methods for quantifying information loss (Duncan, Keller-McNulty and Stokes 2004; Karr et al. 2006). There are many measures because no single one sufficiently captures the true level of utility in the data. The problem is that the actual value of a data file depends on the particular model to be estimated. Even a file with high relative entropy (i.e., a lot of extra noise) may be analytically useful for some models, such as when none of the masked variables is used as a predictor. In other cases, the interest may be only in estimating univariate or bivariate statistics that have been retained by the data masking procedure.

Variance Estimation for Complex Surveys

Geography provides some of the most revealing information about a given subject, and hence public release files typically remove indicators of geographic locations. Unfortunately, this deletion creates particularly acute difficulties for properly analyzing complex survey data. It is common for nationally representative samples to use some geographic unit—counties, blocks of counties, or large metropolian areas—as the primary sampling unit (PSU), but

https://doi.org/10.1017/S1049096511001405 Published online by Cambridge University Press

the use of sampling designs other than simple random sampling (srs) has important implications for estimating standard errors (Wolter 2007). Formulas used to estimate a statistic's variance given srs will not be correct when some combination of stratification and clustering has been used. Methods for correct variance estimation are available in Stata's .svy suite of commands and R's survey library among other software options. However, users must be able to inform the software about which variables in the data file represent PSU and strata information. If such indicators are unavailable, because, for example they have been masked, then the only option is to treat the sample as if all observations were drawn in a single stage with equal probabilities. This, in turn, will cause the software to use incorrect standard error formulas and the analyst to potentially fall victim to Type I or Type II errors.

Worse, it is insufficient to replace explicit PSU and strata identifiers (i.e., names of counties or cities) with seemingly uninformative numeric indicators, as it still may be possible to reconstruct the exact geographic locations by exploiting information implicit in the sampling weights (de Waal and Willenborg 1997; Eltinge 1999). Given these concerns, data collection agencies often pursue a conservative approach in which variables related to the sampling design are suppressed in their entirety. For example, the American National Election Study (ANES) does not release indicators for the strata and sampling units used at each stage of sampling. Lacking variables defining the sampling design, many ANES data users (and similarly masked data files) apply incorrect formulas to estimate the variance of reported statistics.

Thus, the effects of SDC measures are potentially pervasive. Any confidence interval around any statistic estimated on the basis of masked complex survey data may be incorrect. Accurate inference requires specifying the survey design, but there may be prohibitively high hurdles to obtaining that information. The ANES, for example, requires a formal application plus a $335 fee to access its restricted variables. This requirement reflects the high levels of concern that the ANES principal investigators have for protecting the privacy of study participants, and, importantly, almost all commonly analyzed public use surveys take similar steps to mask design variables. Still, design masking places a burden on users to be sensitive to the limitations of the analysis. At present, political scientists usually do not explicitly report whether they use sampling weights or correct formulas for standard errors, so the extent to which users incorrectly assume srs sampling for complex survey data cannot be known.

THE FUTURE OF SDC

Ideally, researchers would have full access to informative data without having to worry about the distortions caused by data masking. Rubin (1993) makes the provocative suggestion that no original observations necessarily need to be released. Instead, researchers could create entirely synthetic datasets that have been imputed based on the relationships observed in the original file. As in the case of multiple imputation for missing data, several different synthetic files could be created to capture the uncertainty in the imputations. Estimation would then be done on each imputed data file separately and combined according to already well-accepted rules for analyzing traditional multiply imputed data (Rubin 1987).

Raghunathan, Reiter, and Rubin (2003) note that there are additional benefits to releasing synthetic data. They argue that, because most surveys are collected according to a complex sampling design, sampling weights and proper variance estimation need to be carefully considered. In contrast, a synthetic file could be created in a form that retains all of the relationships in the original file but whose observations resemble those drawn from simple random sampling. Thus, the end user would not have to be concerned with the typical adjustments that complex survey analysis requires because the sampling design features would be irrelevant for the imputed versions of the datasets.

Despite obvious advantages, very few real-world constructions of entirely synthesized data have been created, and challenges remain. The accuracy of estimation done on entirely imputed data depends on how well the imputation model matches the complicated multivariate process generating the observations. As computational power and statistical theory continue to advance, this approach eventually may be more widely used. Now, however, research is ongoing.

So-called institutional solutions currently provide one realistic, but expensive and inconvenient, route to supply high-utility sensitive data to researchers (National Research Council 2007). Institutional solutions involve a data collecting agency or organization providing access to unmasked data on a highly restricted basis, with options varying substantially in burdens placed on the researcher. At the less restrictive end of the continuum, an agency may require users to sign a restricted use contract before gaining permission to analyze the data on their own computers. At the other extreme, the agency may not allow the researcher to even see the data at all. Instead, the researcher submits code to the agency, and the agency returns the output.

The biggest concern with the institutional approach is that it can be prohibitively expensive for both users and suppliers. Restricted use contracts require administrative support to process the requests and monitor compliance, while limiting data access only to a particular facility requires sufficient hardware, staffing, and building resources. In addition, the user does not have immediate access to the data and, in some instances, must pay an additional fee. These obstacles can limit the use of a data file and, therefore, reduce its scientific impact. The National Opinion Research Center (NORC) and Statistics Netherlands have developed secure *virtual* enclaves that combine remote access with connection restrictions and audit capabilities to minimize these costs. Wider availability of secure virtual enclaves will provide a welcome compromise to the data access/privacy protection tradeoff, and developments in this area continue. Felicia LeClere of NORC is currently investigating the possibility of creating secure, privacy protecting computing instances in the cloud. At the time of this writing, ICPSR—the largest social science data repository in the world—does not offer virtual enclave technology.

CONCLUSION

SDC remains a relatively young field of inquiry, but its importance is likely to grow as the amount of data available to researchers continues to explode. Thus, social scientists need to have a working understanding of the issues involved in protecting the privacy of study participants. This article introduced the field of SDC by defining key terms and describing attempts to quantify risk and utility. The two fundamental messages presented are (1) that disclosure risk may be higher than researchers realize; and (2) the proactive steps data collection organizations take to minimize disclosure risk can affect the ability of the end user to accurately estimate statistical relationships. Research is ongoing to determine solutions that maximize data utility while entirely

https://doi.org/10.1017/S1049096511001405 Published online by Cambridge University Press

eliminating the threat of privacy violations. At this time, however, the only commonly used solutions—short of restricted use contracting—will inevitably affect the quality of inferences. The impact of SDC measures on social science research therefore is potentially pervasive. ■

NOTE

The author thanks Felicia LeClere, Kristine Witkowski, Amy Drayton, Emma Wellstone, William Jacoby, and anonymous reviewers for comments on previous drafts of this article.

REFERENCES

Acquisti, Alessandro, and Ralph Gross. 2009. Predicting Social Security Numbers From Public Data. In *Proceedings of the National Academy of Science*. Vol. 106, 10975–980.

Bethlehem, Jelke G., Wouter J. Keller, and Jeroen Pannekoek. 1990. "Disclosure Control of Micro-data." *Journal of the American Statistical Association* 85 (409): 38–45.

Bialik, Carl. 2010. "Census Bureau Obscured Personal Data—Too Well, Some Say." *Wall Street Journal*, February 6, A2.

Curtis, Andrew J., Jacueline W. Mills and Michael Leitner. 2006. "Spatial Confidentiality and GIS: Re-Engineering Mortality Locations from Published Maps about Hurricane Katrina." *International Journal of Health Geographies* 5: 44–56.

Dalenius, Tore. 1977. "Towards a Methodology for Statistical Disclosure Control." *Statistik Tidskrift* 15: 429–44.

Defays, D., and M. N. Anwar. 1998. "Masking Microdata Using Micro-Aggregation." *Journal of Official Statistics* 14 (4): 449–61.

de Waal, A. G., and L. C. R. J. Willenborg. 1997. "Statistical Disclosure Control and Sampling Weights." *Journal of Official Statistics* 13 (4): 417–34.

Domingo-Ferrer, Josep, and Josep M. Mateo-Sanz. 2002. "Practical Data-Oriented Microaggregation for Statistical Disclosure Control." *IEEE Transactions on Knowledge and Data Engineering* 14 (1): 189–201.

Domingo-Ferrer, Josep, and Vicenc Torra. 2001a. "Disclosure Control Methods and Information Loss for Microdata." In *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, ed. Pat Doyle, Julia J. Lane, Jules J. M. Theeuwes, and Laura M. Zayatz, 91–110. Amsterdam: North Holland.

Domingo-Ferrer, Josep, and Vicenc Torra. 2001b. "A Quantitative Comparison of Disclosure Control Methods for Microdata." In *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, ed. Pat Doyle, Julia J. Lane, Jules J.M. Theeuwes, and Laura M. Zayatz, 111–33. Amsterdam: North Holland.

Doyle, Pat, Julia J. Lane, Jules J. M. Theeuwes, and Laura M. Zayatz, eds. 2001. *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies.* Amsterdam: North Holland.

Duncan, George T., Sallie A. Keller-McNulty, and S. Lynne Stokes. 2004. Database Security and Confidentiality: Examining Disclosure Risk vs. Data Utility through the R-U Confidentiality Map." *Technical Report 142*, Durham, NC: National Institute of Statistical Sciences.

Duncan, George, and Diane Lambert. 1989. "The Risk of Disclosure for Microdata." *Journal of Business and Economic Statistics* 7 (2): 207–17.

Duncan, George T., and Diane Lambert. 1986. "Disclosure-Limited Data Dissemination." *Journal of the American Statistical Association* 81 (393): 10–18.

Duncan, George T., and Sumitra Mukherjee. 2000. "Optimal Disclosure Limitation Strategy in Statistical Databases: Deterring Tracker Attacks through Additive Noise." *Journal of the American Statistical Association* 95 (451): 720–29.

Dwork, Cynthia. 2006. "Differential Privacy." In *Lecture Notes in Computer Science*, Volume 4052, ed. M. Bugliesi, B. Preneel, V. Sassone and I. Wegener, 1–12. Berlin: Springer.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis." In *Lecture Notes in Computer Science*, Volume 3876, ed. S. Halevi and T. Rabin, 265–84. Berlin: Springer.

Eltinge, John L. 1999. "Use of Stratum Mixing to Reduce Primary-Unit-Level Identification Risk in Public-Use Survey Datasets." Washington, DC: Federal Committee on Statistical Methodology Research.

Fellegi, Ivan B., and Alan B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64 (328): 1183–210.

Freese, Jeremy. 2007. "Replication Standards for Quantitative Social Science: Why Not Sociology?" *Sociological Methods and Research* 36 (2): 153–72.

Gomatam, Shanti, and Alan F. Karr. 2003. "Distortion Measures for Categorical Data Swapping." *Technical Report 131*. Durham, NC: National Institute of Statistical Sciences.

Hansell, Saul. 2006. "AOL Removes Search Data on Group of Web Users." *New York Times*, August 8, C4.

Hundepool, A. L., A. L. Willenborg, A. Wessels, L. van Gemerden, S. Tiourine, and C. Hurkens. 1998. *μ-Argus User's Manual*. Voorburg: Statistics Netherlands.

Jaro, Matthew A. 1989. "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida." *Journal of the American Statistical Association* 84 (406): 414–20.

Kargupta, Hillol, and Souptik Datta. 2003. "On the Privacy Preserving Properties of Random Data Perturbation Techniques." http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.6457

Karr, A. F., C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. 2006. "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality." *Technical Report 153*. Durham, NC: National Institute of Statistical Sciences.

King, Gary. 1995. "Replication, Replication." *PS: Political Science and Politics* 28 (3): 444–52.

Lambert, Diane. 1993. "Measures of Disclosure Risk and Harm." *Journal of Official Statistics* 9 (2): 313–31.

Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. 2007. "t-Closeness: Privacy beyond *k*-Anonymity and ℓ-Diversity." In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, Istanbul, 106–15.

Machanavajjhala, Ashwin, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubra-maniam. 2006. "ℓ-diversity: Privacy beyond *k*-Anonymity." In *International Conference on Data Engineering*.

McDermott, Rose. 2010. "Data Collection and Collaboration." *PS: Political Science and Politics* 43 (1): 15–16.

Muralidhar, Krish, and Rathindra Sarathy. 2010. "Does Differential Privacy Protect Terry Gross' Privacy?" Paper available at http://gatton.uky.edu/faculty/muralidhar/PSD2010B.pdf. Accessed June 20, 2011.

Narayanan, Arvind, and Vitaly Shmatikov. 2008. "Robust De-Anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)." In *IEEE Symposium on Security and Privacy*, 111–25.

National Research Council. 2007. *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*. Washington, DC: The National Academies Press.

Paaß, Gerhard. 1988. "Disclosure Risk and Disclosure Avoidance for Microdata." *Journal of Business and Economic Statistics* 6 (4): 487–500.

Raghunathan, T. E., J. P. Reiter, and D. B. Rubin. 2003. "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics* 19 (1): 1–16.

Reiss, Steven B. 1984. "Practical Data-Swapping: The First Steps." *ACM Transactions on Database Systems* 9 (1): 20–37.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponses in Surveys*. New York: John Wiley and Sons.

Rubin, Donald B. 1993. "Discussion: Statistical Disclosure Limitation." *Journal of Official Statistics* 9 (2): 461–68.

Sweeney, Latanya. 1997. "Weaving Technology and Policy Together to Maintain Confidentiality." *Journal of Law, Medicine, and Ethics* 25 (2–3): 98–110.

Sweeney, Latanya. 2000. *Uniqueness of Simple Demographics in the US Population*. Technical report. Pittsburgh, PA: Carnegie Mellon University.

Sweeney, Latanya. 2002. "*k*-anonymity: A Model for Protecting Privacy." *International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems* 10 (5): 557–70.

Willenborg, Leon, and Ton de Waal. 2001. *Elements of Statistical Disclosure Control*. New York: Springer-Verlag, Inc.

Wolter, Kirk M. 2007. *Introduction to Variance Estimation*. Second ed. New York: Springer.

Xiao, Xiaokui, and Yufei Tao. 2007. "m-invariance: Towards Privacy Preserving Re-Publication of Dynamic Datases." In *SIGMOD: Special Interest Group* on Management of Data. Beijing, China. June 11–14.

Zayatz, Laura. 2006. *Disclosure Avoidance Practices and Research at the US Census Bureau: An Update*. Technical report. Washington, DC: US Census Bureau.

https://doi.org/10.1017/S1049096511001405 Published online by Cambridge University Press