Optimizing the Use of Micro-data: An Overview of the Issues Presented at I Quality and Access to Federal Data: Memorial Session in Honor of Pat J. Doyle

Julia Lane[i]
Senior Vice President and Director, Economics, Labor and Population
National Opinion Research Center, University of Chicago, 55 East Monroe Street, Chicago, IL 60603
Lane-julia@norc.uchicago.edu

Abstract

New capacities to collect and integrate data offer expanded potential for scientists and policy-makers to understand factors contributing to key national priorities. However, two substantial challenges face collectors and producers of economic data as a result of this increased capacity. The first is how can the information derived from vast streams of data on human beings be used while protecting confidentiality? The second is the essence of good science: how can society best provide and promote access to rich and sensitive data so that empirical results can be generalized and replicated? This paper begins by discussing current confidentiality protection techniques accompanied by illustrations of some consequences for the typical type of analyses performed by economists. It then describes the challenges that are emerging as a result of technological advances, and develops a simple economic framework. The paper concludes with a suggested research agenda.

Keywords: confidentiality, micro-data access, cyberinfrastructure

" It is becoming clear that advances in technology and increased use of administrative records may, at some point in the future, render our current disclosure avoidance procedures inadequate. At the same time the… federal statistical system face[s] increasing demands for more, better and more recent data to meet critically important public policy and research needs."[ii]

"The extraordinary growth of electronic infrastructure, capacity, and use in the past decade has posed a profound new set of questions about the control, dissemination, power and use of information. On the one hand the high speed internet and the World Wide Web, email, electronic shopping, and cell phone use have opened up extraordinary new worlds of communication and are changing the way we work, play, and learn. On the other, as the electronic world enters our daily lives, the private space untouched by the intrusions of cyberspace and information seekers shrinks - for individuals, firms, and organizations. …There is also another challenge. The need to build more efficient surveillance networks to combat potential terrorist attack argues for less privacy for the individual person or firm to guarantee the security of the society in general. It is in this environment that citizens, business and technology leaders, and policy makers have to figure out how to understand, manage, and regulate the new cyberworld."[iii]

1. Introduction

New capacities to collect and integrate data offer expanded potential for scientists and policy-makers to understand factors contributing to key national priorities –like job, income and wealth creation, as well career path and retirement decisions made by individuals. This capacity can also contribute to meeting a critical national security need. The major security threat to the United States is inherently human and an improved ability to understand and predict malevolent behaviors can provide one means for addressing that threat.

Two substantial challenges face collectors and producers of economic data as a result of this increased capacity. The first is how can the information derived from vast streams of data on human beings be used while protecting confidentiality? The second is the essence of good science: how can society best provide and promote access to rich and sensitive data so that empirical results can be generalized and replicated?

An existing community has already focused on protecting confidentiality. In particular, federal statistical agencies have devoted substantial resources to both statistical and technical ways to protect confidentiality[iv], the Social and Behavioral Research Working Group recently drafted a report entitled "Achieving Effective Human Subjects Protection and Rigorous Social and Behavioral Research" for the

Human Subjects Research Subcommittee of the Committee on Science, National Science and Technology Council, PITAC[v] recently issued a report on cybersecurity that addressed some confidentiality issues, and numerous studies have been undertaken by the National Academy of Sciences and the Committee on National Statistics. The National Science Foundation has also been active in the area of cybertrust, and the PORTIA (Privacy, Obligation and Rights in Technologies of Information Assessment) project based at both Yale and Stanford universities directly addresses many of the key issues.

However, focusing on confidentiality protection alone will lead to piecemeal approaches and result in outcomes that are neither in the best interests of decision-makers nor of society at large. The appropriate approach is to optimize the amount of data access, subject to meeting key confidentiality constraints. And, although Fienberg and Duncan (2003, 2004), in particular, have been vocal advocates of preserving statistical utility of tabular data, little attention has been paid to optimizing access to micro-data.

This paper begins by discussing current confidentiality protection techniques accompanied by illustrations of some consequences for the typical type of analyses performed by economists. It then describes the challenges that are emerging as a result of technological advances, and develops a simple economic framework. The paper concludes with a suggested research agenda.

2. An Overview Of Current Confidentiality Protection Techniques and Their Consequences

A good description of the practical application of micro-data disclosure limitation techniques practiced at the U.S. Census Bureau is provided in Zayatz (2005). She points out that the risk of reidentification can be reduced either by reducing the amount of information or by perturbing the data

The means used to reducing the amount of information include variable deletion, recoding categorical variables into larger categories (perhaps using thresholds), recoding continuous variables into categories, rounding continuous variables, using top and bottom codes, using local suppression and enlarging geographic areas. Data can be perturbed by means of noise addition, record swapping, rank swapping, blanking and imputation, micro-aggregation or by multiple imputation/modeling to generate synthetic data.

Although each of these approaches can have an impact on the validity of social science analysis, the decision to apply them is made independently of the potential consequences. A good discussion of the issues is provided in Smith (1991). The impact of decisions on topcoding is well summarized in examining the earnings inequality literature. Burkhauser et al. (2004) find that changes in one of the most important public use surveys, the Current Population Survey, topcoding rules in the 1990's artificially *increased* measured earnings inequality.

The key problem is that a standard measure for calculating earnings inequality is the Gini coefficient which ranges between 0 and 1. A value of 0 corresponds to a situation where everyone has the same income, or perfect equality. The value of the coefficient increases as the richest percentiles in society earn higher proportions of income. Topcoding artificially reduces the maximum income level, resulting in a coefficient that is biased down. Arbitrary changes in topcodes can change the Gini coefficient up or down – artificially changing earnings inequality.

As Mishal and Bernstein point out in a debate between Robert Lerman (1997) and Lawrence Mishal/Jared Bernstein (1997).

> "However, before we can reliably measure inequality trends in the CPS or, for that matter, any other public-use data set, we must deal with the issue of top codes, an issue that becomes particularly germane when earnings at the top are growing quickly relative to those elsewhere in the earnings distribution. …There are a number of ways to approach the top-coding problem. One is simply to ignore top coding. Doing this, however, is a problem in Gini analysis, because nominal wage growth over a period when the top code does not change or increases only slightly will lead to increasing shares of earners who are top coded, thus biasing the Gini coefficients downward. Such a downward bias applied between 1981 and 1987, when the top code stayed between $75,000 and $99,999, before doubling in 1988." pp 3-4.

A visual illustration of the consequences are clear from the graph reproduced from Burkhauser et al. (2004) below. The bottom line in the graph shows that had topcoding on the Current Population Survey been consistent, then earnings inequality, as captured by the Gini coefficient, would have increased steadily between 1975 and 2001. However, topcoding did not remain consistent. The second line plots the Gini

coefficient derived from public use files. The public-use topcode was $99,999 until 1995 when the Census Bureau both raised the public-use topcode to $150,000 and assigned cell means for persons with earnings above the topcode. The surge in earnings inequality from about .34 to .39 is completely an artifact of that topcoding decision. It is worth noting that the 1993 surge in the third line reflects a data collection, rather than a reporting decision. In that year, the Census Bureau changed its internal system to permit the recording of incomes of $999,999, rather than $249,000 (between 1979 and 1984, the maximum permissible was $99,999).

In sum, the approach used to topcoding public use data can result in vastly different information being provided to policy makers. And, to repeat, inappropriate action by the policy makers can result in outcomes that are neither in the best interests of decision-makers nor of society at large.

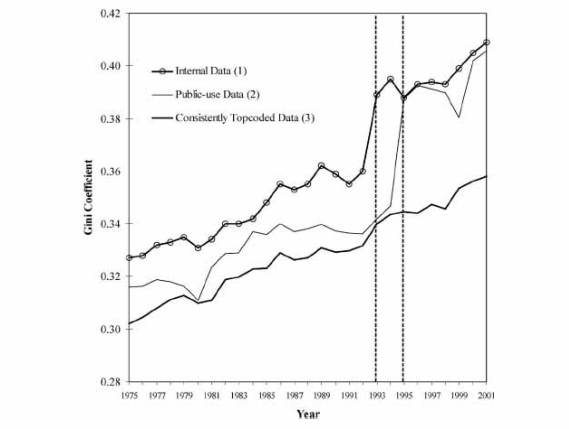

Fig. 1. Trends in Gini coefficients of earnings for full-time year-round workers (1975–2001). (1) Jones and Weinberg (2000), Table 1, page 3 and US Census Bureau (2002). (2 and 3) Authors' calculations using CPS public-use data, 1976–2000.

The consequences of topcoding on other standard uses of public use files are clear, since the theory associated with regressions when the dependent variables is censored from both above and below is well developed. Indeed, the 2000 Nobel Prize was given, in part, to Jim Heckman for his pathbreaking work on statistical approaches to deal with the econometric problems posed by selective samples.

A brief example using an earnings regression model illustrates the effect on regression coefficients. Suppose we have an earnings regression model
$Y_i = X_i'\beta + \varepsilon_i \quad \varepsilon_i \sim N(0,\sigma^2)$.
Where Y_i is the earnings of individual i, and X_i is a set of that individual's characteristics, but the model is censored from below by a and above by b. It is straightforward to show that standard least squares regression will result in slope coefficients that are downwardly biased. Consistent coefficients can be estimated if the distribution of the error term given the regressors is known, and some of Heckman's most important work has dealt with doing just this by modeling the behavioral decision that leads to censoring from below. The fundamental problem with arbitrary topcoding is that the distribution is not provided, making it extremely difficult to recover consistent estimates. Although several alternative estimators have been developed, using different assumptions about the distribution underlying the topcoded values, there is still wide divergence among estimated coefficients.

The implications of this divergence can be quite substantial. Two key issues of interest to policymakers are the black/white earnings gap and the return to education. The following Table, which is reproduced from Chay and Powell (2001), illustrates the wide divergence in estimates using different techniques (using topcoded Social Security data). The first two columns (OLS1 and OLS2) use ordinary least squares approaches that do not attempt to address the distributional consequences of topcoding. Maximum Likelihood Estimation procedures (MLE), the results of which are reported in column 3, assume that the errors normally distributed and homoskedastic. Chay and Powell develop three semiparametric estimators for censored regressions: the CLAD (censored least absolute deviations), symmetrically censored least squares (SCLS) estimation method, and the identically censored least absolute deviations (ICLAD).

The first panel reflects the results of running standard earnings regressions that estimate the effect of race on the log of earnings using these six different approaches. The coefficients reported in each column can be approximately interpreted as the percentage difference in earnings between blacks and whites in each year, controlling for age. Briefly, not only do estimates of the black/white earnings gap range from .35 to .63 log points in 1963, but estimates of the degree to which the gap closed between 1963 and 1971 range from .06 log points in black/white earnings gap using OLS regression techniques to .15 log points using alternative measures. Policy makers might look at one set of numbers and conclude that the racial earnings gap was closing rapidly; at another set and conclude that it was closing slowly. In the former case, the policymaker might well conclude that no intervention was required; in the latter, that intervention was necessary. One of those decisions would be wrong, although it is not clear which is the incorrect decision. Certainly one would be neither in

the best interests of decision-makers nor of society at large.

The second panel reflects the results of using the different estimation techniques to calculate the return to education – another topic of key interest to policymakers. A policy maker who only used information from the second column would note that the returns to education had gone from about 1% in 1963 to approximately zero in 1973, and would be forgiven for concluding that further investment in education was unnecessary. A policymaker examining the final column would see that the return to education was a consistent 7%, and could conclude that further investment would be a wise allocation of public monies.

Estimated Effects of Race and Education on Log-Earnings
(estimated standard errors in parentheses)

	OLS1	MLE	CLAD	SCLS	KLAD
Black-White Gap					
1963	-0.355	-0.629	-0.416	-0.444	-0.474
	(0.033)	(0.044)	(0.027)	(0.031)	(0.032)
1971	-0.242	-0.486	-0.244	-0.287	-0.312
	(0.031)	(0.044)	(0.022)	(0.032)	(0.031)
Returns to Education					
1963	0.041	0.102	0.051	0.068	0.073
	(0.003)	(0.004)	(0.004)	(0.007)	(0.003)
1971	0.035	0.100	0.054	0.065	0.070
	(0.002)	(0.004)	(0.003)	(0.005)	(0.003)

Notes: The dependent variable is the natural logarithm of annual taxable earnings. Regressions also include a constant, and age and age-squared as explanatory variables. Observations with non-positive earnings are dropped from the analysis. The sample sizes for 1963, 1964, 1970, and 1971 are 8525, 8529, 8391, and 8275, respectively. The OLS2 specification also drops top-coded observations, leading to sample sizes of 4632, 4267, 4485, and 4163. MLE is Tobit maximum likelihood; CLAD is censored least absolute deviations; SCLS is symmetrically censored least squares; ICLAD is identically censored least absolute deviations.
Source: Adapted fromChay and Powell (2001)

The consequences of the other disclosure limitation techniques – such as recoding, rounding and dataswapping – are less well documented, although each should act to bias coefficients towards zero. It is remarkable, however, that despite the fact that statistical agencies publish extensive and high quality documentation that inform users of the consequences different sampling procedures and nonsampling errors, and how to adjust estimates accordingly[vi], there is no comparable effort for disclosure limitation. It would seem obvious that the holder and producer of microdata should list specific limitations that affect the ability of the microdata to support valid analyses. Alternatively, the data producer should either provide access to suitable microdata so that users can determine which types of estimation procedures to use,, or provide suitable auxiliary information with public use microdata so to permit the approximate reproduction of the results that might be obtained on the original microdata.

Part of the challenge is that social scientists use micro-data in many different ways and it is difficult to directly define what is meant by data quality. An illustrative example is the workshop on total survey error that the National Institute of Statistical Sciences (NISS) held in March 2005, from which it is clear that quality concepts are difficult to use in most specific settings [vii]. The Eurostat definitions, which lack metrics, are (1) relevance, (2) accuracy, (3) timeliness, (4) accessibility and clarity of results, (5) comparability, (6) coherence, and (7) completeness (Haworth et al. 2001). Winkler (2005e) has provided some metrics to diagnose serious problems with a file, but these do not assure analytic quality. As Winkler (2005f) has pointed out, the challenge in maintaining quality in a masked file is due to the fact that certain aggregates such as higher order moments must be accurate (say for regressions).

3. Future Data Collections and the Associated Confidentiality Challenges

The previous section demonstrated that current statistical disclosure techniques act in unknown ways to severely diminish the utility of micro-data for analysis. It is also clear that the challenge to protecting the confidentiality of micro-data will only increase. In addition to the challenges posed by the increased capacity for reidentification, that were thoroughly documented in Doyle et al. (2001), new data collection modalities are emerging that pose much greater likelihood of reidentification, and there is much greater access to administrative data.

Although data collection on individuals and organizations has historically consisted of either survey based or administrative data, cyberinfrastructure [viii] advances have fundamentally changed the way in which scientists are collecting information and modeling human behavior. Indeed, a recent National Science Foundation solicitation, entitled “Next Generation Cybertools” noted that new ways have been developed to improve both domain-

specific and general-purpose tools to analyze and visualize scientific data -- such as improving processing power, enhanced interoperability of data from different sources, data mining, data integration, information indexing[ix]. And a calculation at the recent NSF supported workshop[x] about how many terabytes of data would be necessary to capture an entire life on video found that if the life were recorded on low web video, at 50 kbits/sec, the total space required would be 15TB. Even with DVD quality recording, t 5Mbits/sec, the total storage would be 1500TB . Clearly, an entire life can now be captured and stored on existing media.

In addition, while academic social scientists are increasingly using these cybertools to combine data from a variety of sources -- including text, video images, wireless network embedded devices and increasingly sophisticated phones, RFID's[xi], sensor webs, smart dust and cognitive neuroimaging records, the same is also true for the private sector.

> "Workers in warehouses across Britain are being "electronically tagged" by being asked to wear small computers to cut costs and increase the efficient delivery of goods and food to supermarkets, a report revealed yesterday… Under the system workers are asked to wear computers on their wrists, arms and fingers, and in some cases to put on a vest containing a computer which instructs them where to go to collect goods from warehouse shelves. The system also allows supermarkets direct access to the individual's computer so orders can be beamed from the store. The computer can also check on whether workers are taking unauthorised breaks and work out the shortest time a worker needs to complete a job." Hencke, The Guardian, 2005[xii]

The capacity for this new technology to push forward the frontiers of social science research and answer important societal questions is clear. However, the progress will also put substantial pressure on statistical agencies to create and provide access to such data in order to keep pace with the private sector. Obvious new confidentiality challenges arise with these advances – such as protecting the identity of individual video images.

In addition to new data collection modalities, advances in cyberinfrastructure also mean that much more administrative data can be stored and disseminated. Census Bureau research has shown that the wide availability of certain kinds of personal information increases the chance of disclosure of confidential information—particularly when date of birth and geography are available. However, there is increasing open access to state level administrative records that people can use to identify respondents such as birth records and marriage records[xiii]. These records can be combined with other sources to increase the risk of reidentification.

Although there have been substantial advances in statistical disclosure protection techniques (see, for example, Winkler, 2005, and some of the ideas put forward at a recent workshop organized by Dwork and Fienberg, 2005) in response to some of these reidentification threats, little of this has been accompanied by a discussion of the impact on data quality, although Kaufmann et al (2005) do discuss the impact of masked procedures on data quality for a particular survey. This lack of attention is a major threat to high quality empirical social science research, given the degradation in quality stemming from the use of current disclosure protection techniques and documented in the previous section.

4. An Economic Framework

Putting the challenge to the statistical system in an economic framework, the data custodian is charged with maximizing data utility subject to both cost and reidentification constraints[xiv]. Each of these is discussed in more detail below.

There is a full discussion of the utility of micro-data in Lane (2003). Assume that the mission of each statistical agency is to maximize the utility to society, conditional on keeping disclosure risk at a predefined level. Define U as data utility, the value to society of micro-data access. This utility depends on a number of factors, data quality, researcher quality, and the number of times the data are accessed. Let Q= Data quality, R=Researcher quality, and N=number of times the data are accessed. Then we have U= u(Q, R, N),
Data quality depends on the portfolio of access modalities available to the research community. If M_i = modality i, then we can write $Q(M_i)$. R and N are both determined by the access costs, A, imposed by the access modality, and we can therefore write R and N as functions of A: $R(A_i)$ and $N(A_i)$.

The expected costs to society of micro-data access can be defined as the harm to individuals or organizations should disclosure occur, H, times the probability of disclosure, D, plus the monetary cost of providing access, C . The probability of disclosure is typically set at a "target" level: since most agencies are charged with using reasonable means to protect data, this implicitly means setting re-identification risk to some fixed number. Thus, the expected social cost, S, can be written as $S = H \cdot D + C$

The factors contributing to the target risk of disclosure D^* can be written as $D^* = d(E, I, Z, M_i)$
where E is the existence and accessibility of other data sources that can be used for reidentification. The relationship between this and re-identification is affected by technology, T, and can be written E(T)
I is the existence of malevolent interlopers. This relationship is affected by technology, legal penalties, L, and the characteristics of the population, X and can be written I(T, L, X). Z is researcher error. This is affected by technology, legal penalties, training and adoptable protocols, P and can be written Z(T,L, P). M, as before, is the set of access modalities. Harm, H, is also likely to be a function of population characteristics, and can be written H(X). Finally, the monetary cost constraint is $C = p_t T + \Sigma_{Mi} p_{Ai}M_i$ where p_i reflects the price of providing a certain level of protection.
The constrained optimization is then to maximize utility subject to the constraint $S - C - H D^* \leq 0$, or maximize the associated Lagrangian $L = U - \lambda (H d(E,I,Z, M_i) + p_t T + \Sigma_{Mi} p_{Ai}M_i - S)$ In general, maximization requires that the marginal benefits with respect to each variable are set equal to the marginal costs. This, in turn, means that the statistical agency needs to be able to quantify the relative marginal value of each of the key input variables, which is no trivial task. The following section offers some suggestions towards this goal.

5. Using the Framework To Shape a Research Agenda

This framework, despite the somewhat cumbersome notation serves the important function of identifying key focus areas for confidentiality research – namely:

1. *Developing metrics of data quality* **Q**

The work of Domingo Ferrer and Torra (2001) and Duncan et al. (2001) which attempted to quantify information loss at the same time as measuring disclosure risk began to outline an approach that should be further advanced. And Shlomo (2005) proposed a series of measures for frequency tables such as dDistance metrics to measure distortions to distribution and expected totals, non-parametric statistical testing for same location, scale and shape of empirical distributions, Impact on statistical inference, such as: Variance of cell size, Chi-Square measures of association, Pearson and log-likelihood ratio testing for log-linear modelling, and "Between" variation of target variables as expressed by R2.

However, similar metrics for micro-data have not been developed. The two examples provided in section 2 are illustrative of the issues in that measures of inequality require knowledge of the entire distribution; accurate measurement of key coefficients requires knowledge of the relationship among variables. This point has also been made by Winkler (2005b, c and d) who notes the importance of developing measures that reflect the specific analytic use of the files.

Multiple approaches could be taken to determine these uses – for example a literature review that summarized the main uses for major public use data sets; another to survey key federal and academic users.

2. *Quantifying the effect of the cost of access A on usage N and researcher quality R*

The work by Dunne (2001) and Seastrom (2001) outlined some of the key issues associated with imposing high costs to researcher access. In the NSF award that served as one of the forces initiating the Longitudinal Employer-Household Dynamics program, Abowd, Haltiwanger and Lane (1998) pointed out that for more than two decades, public policy around the world was influenced by analysis of public-use American micro-data samples. However, the increasing availability of administrative data, as well as data from other countries, combined with the cost (including the cost of time) of accessing U.S. federal data now means that many of the best researchers in the country, and in the world, have found alternative datasources for their empirical analysis.

Quantifying the effect of the cost of access, and using this as a basis for informed decision making would clearly be difficult. However, one possible approach would be to survey ten years of the relevant academic and federal literature and document how often federal data are used as a basis for analysis, relative to other sources, as well as identify any trends. Similarly, a survey of top federal and academic researchers would help identify the relationship between access and use.

3. *Measuring harm* **H**

Madsen (2003) outlined many of the key philosophical issues in an NSF workshop held in 2003 [xv]. He identified a key privacy paradox as follows:

> The "privacy paradox" occurs when data managers interpret the right to privacy as a near absolute ethical standard. Such an understanding of the nature of the right to privacy leads to an extreme understanding of the nature of the responsibility of confidentiality with newer and more restrictive controls on data access. More privacy in the research context paradoxically

results in less social benefit, rather than in more" (p3).

Researchers such as Singer (2001) and Greenia et al. (2001) have attempted to quantify harm, but an extensive research agenda remains. Both Greenia and Singer have since noted that the research agenda has also substantially changed since the events of September 11, 2001 both because government data collection activities have increased and because public perception of the harm associated with such collection is likely to have changed.

4. *Quantifying the relationship between other data sources **E** and disclosure **D***

Winkler (2003a and b, 2004 1,b,c and d and 2005a) as well as Domingo-Ferrer and Torra (2001, 2003) have outlined an extensive research agenda.

5. *Modelling malevolent behavior **I** and researcher error **Z***

A recent NSF workshop on cyberinfrastructure and the social sciences included, as one theme, the importance of using social science to understand and model malevolent behavior [xvi] . As was pointed out, the importance of this goes far beyond the federal statistical community, since such behavior affects a wide variety of realms – ranging from financial and personal harm (Data and money, identity theft) to cyber-terrorism,
'Phishing' and 'pharming', denial of service attacks, hacktivism, hate crimes, and to gambling and pornography. The summary report (see Berman and Brady, 2005), noted that in this area:

> "Social scientists can be especially helpful in developing an understanding of the motivations and capacities of those who might engage in malevolent behavior, in designing institutions and procedures that deter malevolent behavior and that produce trustworthy Cyberinfrastructure."

Indeed, there is a group of researchers – such as Joan Feigenbaum and Deb Agarwal – that has established a strong knowledge base in trust management issues and collaborative computing environments. Salvatore Stolfo and Roy Maxion have similarly extensive research agendas to detect data mining based intrusion and to develop behavior based computer security models[xvii] [20]

Hence, a sensible research agenda for the statistical community might well be to join forces with researchers to better model malevolent behavior, and develop sensible deterrents. The corollary would be to combine resources with other federal and private institutions that have common concerns.

6. *Investigating alternative technological approaches **T** to providing new access modalities **M***

Protecting databases against intruders has a long history in computer science (a classic article is Dobkin, Jones and Lipton, 1979). Computer scientists themselves are interested in protection the confidentiality of the data on which they do research (for example, the Abilene Observatory supports the collection and dissemination of network data, such as IP addresses). [xviii] Cyberinfrastructure advances have certainly served to expand the set of access modalities, particularly with respect to remote access. The cybertrust initiative at NSF has created an entire research community that focuses on creating network computers that are more predictable and less vulnerable to attack and abuse, that is developed, configured, operated and evaluated by a well-trained workforce, and that educates the public in the secure and ethical operation of such computers. The Department of Defense has developed different levels of web-based access ranging from unclassified (nipr-net) to secret (sipr-net) to top-secret (jwics-net)[xix] using off the shelf technology. Similarly, the PORTIA project focuses on both the technical challenges of handling sensitive data and the policy and legal issues facing data subjects, data owners and data users. Finally, the recent NSF SBE/CISE workshop on cyberinfrastructure [xx] outlined a combined computer and social science research agenda for different approaches to access.

In addition, several agencies have preexisting institutional structures that could be used to expand the number and types of access modalities: such as the Census Bureau's Research Data Centers and the data enclave at NCHS. Similarly, the National Science Foundation funded supercomputer centers could be deployed to provide a portal for information about advances in confidentiality research, provide training about confidentiality procedures to researchers and institutional review boards, as well as provide computational facilities to develop both technical and non-technical solutions to confidentiality problems. Finally, the European Union is also making a substantial investment in a centralized location for social science data, and in the associated confidentiality issues, as part of its VIIth Framework.

6. Summary

Economists should act to promote the view that the federal statistical agencies, and other data custodians, be as concerned about providing data to their customers and about promoting use of their data as they is about protecting their respondents and ensuring the security of confidential information. The activities needed to avoid what some have called a pending "train wreck" between respondents, data custodians and data users involve technological advances, legal strategies, policy enhancements (related to both privacy and disclosure avoidance in the context of survey and census data as well as in the context of administrative data), interagency coordination, new disclosure avoidance techniques, and privacy research.

This paper has attempted to formalize a number of the issues and ideas that have circulated in disparate arenas. It began by noting that the study of confidentiality remains quite piecemeal in nature, without an overarching framework to provide context. It highlighted the particular problems posed by a pursuit of confidentiality protection that did not pay attention to the main aim of providing data access, namely data utility, arguing that this could distort information and potentially lead to incorrect decisions. It outlined a standard economic approach to thinking about the optimization problem, provided a brief list of new initiatives and outlined a possible research agenda for optimizing access to micro-data.

References

Abowd, J, J. Haltiwanger, and J. Lane "Dynamic Employer-Household Data and the Social Data Infrastructure", National Science Foundation, SES-9978093, September 28, 1999 - September 27, 2003

Abowd, J and J Lane "The Economics of Data Confidentiality", mimeo, Committee on National Statistics 2003 www7.nationalacademies.org/cnstat/Abowd_Lane.pdf

Berman, F and H. Brady "Final Report: NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences", May 2005, available at www.sdsc.edu/sbe/.

Bernstein, J. and Mishal, L "Has Earnings Inequality Stopped Growing?" Monthly Labor Review, December 1997, 3-16

Burkhauser, R, J. Butler, S. Feng and A. Houtenville "Long term trends in earnings inequality: what the CPS can tell us" Economics Letters, 82(2), February 2004, 295-299.

Chay, K and J. Powell "Semiparametric Censored Regression Models" Journal of Economic Perspectives" 15(4), Fall 2001, 29-42

Dobkin, D., A. Jones and R. Lipton, "Secure Databases: Protection Against User Influence" ACM Transactions on Database Systems (TODS) Volume 4 , Issue 1 (March 1979) Pages: 97 - 106

Domingo-Ferrer, J., V. Torra, Disclosure control methods and information loss for micro-data, in: P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.), Confidentiality, Disclosure and Data Access, North-Holland, Amsterdam, 2001, pp. 91-110.

Domingo-Ferrer, J., V. Torra, A quantitative comparison of disclosure control methods for micro-data, in: P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.), Confidentiality, Disclosure and Data Access, North-Holland, Amsterdam, 2001, pp. 111-133.

Domingo-Ferrer, J and V Torra "Advanced Record Linkage for Disclosure Risk Assessment" 2003, mimeo, presented at NSF workshop on confidentiality.

Duncan, G, S.E. Fienberg, R. Krishnan, R. Padman, S.F. Roehrig, Disclosure limitation methods and information loss for tabular data, in: P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.), Confidentiality, Disclosure and Data Access, North-Holland, Amsterdam, 2001, pp. 135-166

Duncan, G. T., Keller-McNulty, S., & Stokes, S. L. Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report 2003-6. Heinz School of Public Policy and Management, Carnegie Mellon University. 2003

Duncan, G Exploring the Tension Between Privacy and the Social Benefits of Govermental Databases, mimeo, Carnegie Mellon University, 2004

Dunne, T. (2001). Issues in the establishment and management of secure research sites. in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, eds Pat Doyle, Julia Lane, Laura Zayatz and Jules Theeuwes, North Holland, 2001.

Greenia, N, J.B. Jensen and J. Lane., "Business Perceptions of Confidentiality." in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, eds Pat Doyle, Julia Lane, Laura Zayatz and Jules Theeuwes, North Holland, 2001.

Haworth, M., Bergdahl, M., Booleman, M., Jones, T., and Magaleno, M. (2001). "LEG chapter on

Quality Framework," Proceedings of Q2001, Stockholm, Sweden, May 2001, CD-ROM.

Hencke, D "Firms tag workers to improve efficiency" Tuesday June 7, 2005, The Guardian

Kaufman, S., Seastrom, M., and Roey, S. (2005), "Do Disclosure Controls to Protect Confidentiality Degrade the Quality of the Data?" *American Statistical Association, Proceedings of the Survey Research Section*, to appear.

Lane, J "The Uses of Micro-data" Keynote speech to Conference of European Statisticians, Geneva, Switzerland, 2003 http://www.unece.org/stats/documents/ces/2003/crp.2.e.pdf

Lane, J "Key Issues in Confidentiality Research: Results of an NSF workshop" http://www.nsf.gov/sbe/ses/mms/nsfworkshop_summary1.pdf May 2003

Lerman, R "Reassessing trends in earnings inequality" Monthly Labor Review, December 1997, 17-25

Madsen, P. "The Ethics of Confidentiality: The Tension Between Confidentiality and the Integrity of Data Analysis in Social Science Research", mimeo, Carnegie Mellon University, June 2003

Rodgers, W., C. Brown and G. Duncan "Errors in Survey Reports of Earnings, Hours Worked, and Hourly Wages" Journal of the American Statistical Association, 88(424) December 1993, 1208-1218

Seastrom, M. "Licensing." 2001. Pages 341-370 in Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. (Editors: Pat Doyle, Julia Lane, Jules Theeuwes, and Laura Zayatz). New York: Elsevier

Shlomo, N "Information Loss Measures For Frequency Tables", mimeo, Southampton Statistical Sciences Research Institute, 2005.

Singer, Eleanor. "Public Perceptions of Confidentiality and Attitudes Toward Data Sharing by Federal Agencies." 2001 in Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. (Editors: Pat Doyle, Julia Lane, Jules Theeuwes, and Laura Zayatz). New York: Elsevier

Smith, J "Data Confidentiality: A Researcher's Perspective" Proceedings of the American Statistical Association, Social Statistics Section, pp. 117-120, 1991

U.S. Department of Labor, Current Population Survey Design and Methodology, Technical Paper TP63RV, March 2002, Washington DC.

Winkler, William "Micro-data Confidentiality References", mimeo, US Census Bureau, 11 February 2005.

Winkler, W. E. (2003a), "Methods for Evaluating and Creating Data Quality," Proceedings of the ICDT Workshop on Cooperative Information Systems, Sienna, Italy, January 2003.

Winkler, W.E. (2004a), "Re-identification Methods for Masked Micro-data," in (J. Domingo-Ferrer and V. Torra, eds.) Privacy in Statistical Databases 2004, New York: Springer, 216-230.

Winkler, W.E. (2004b), "Masking and Re-identification Methods for Public-Use Micro-data: Overview and Research Problems," in (J. Domingo-Ferrer and V. Torra, eds.) Privacy in Statistical Databases 2004, New York: Springer, 231-247, http://www.census.gov/srd/papers/pdf/rrs2004-06.pdf .

Winkler, W. E. (2004c), "Record Linkage: Overview of Recent Developments and Applications," in (S. Biffignandi, ed.) Combining Data from Different Sources – Applications of Record Linkage Methodology and Estimation Using Administrative Data, Rome: ISTAT, to appear.

Winkler, W. E. (2005a), "Overview of Record Linkage and Current Research Directions," U.S. Bureau of the Census, Statistical Research Division Report at http://www.census.gov/srd/, to appear.

Winkler, W. E. (2005b), "Data Quality in Data Warehouses," in (J. Wang, ed.) Encyclopedia of Data Warehousing and Data Mining, to appear.

Winkler, W. E. (2005c), "Methods and Analyses for Determining Quality," to appear.

Winkler, W. E. (2005d), "Data Quality for Modeling, Analysis, and Data Mining," to be submitted.

Winkler, W. E. (2005e), "Methods and Analyses for Determining Quality," 2nd Keynote address at the 2005 *ACM SIGMOD Workshop on Information Quality in Information Systems* (available under Post Workshop Material at http://iqis.irisa.fr/).

Winkler, W. E. (2005f), "Modeling and Quality of Masked Microdata," *American Statistical Association, Proceedings of the Survey Research Section*, to appear.

Zayatz, L. (2005), "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update", Research Report Series (Statistics #2005-06), Statistical Research Division, U.S. Census Bureau, Washington, D.C.

atz). New York: Elsevier

Smith, J "Data Confidentiality: A Researcher's Perspective" Proceedings of the American

Statistical Association, Social Statistics Section, pp. 117-120, 1991

U.S. Department of Labor, Current Population Survey Design and Methodology, Technical Paper TP63RV, March 2002, Washington DC.

Winkler, William "Micro-data Confidentiality References", mimeo, US Census Bureau, 11 February 2005.

Winkler, W. E. (2003a), "Methods for Evaluating and Creating Data Quality," Proceedings of the ICDT Workshop on Cooperative Information Systems, Sienna, Italy, January 2003, longer version in Information Systems (2004), 29 (7), 531-550.

Winkler, W. E. (2003b), "Data Cleaning Methods," Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification, Washington, DC, August 2003.

Winkler, W.E. (2004a), "Re-identification Methods for Masked Micro-data," in (J. Domingo-Ferrer and V. Torra, eds.) Privacy in Statistical Databases 2004, New York: Springer, 216-230, http://www.census.gov/srd/papers/pdf/rrs2004-03.pdf .

Winkler, W.E. (2004b), "Masking and Re-identification Methods for Public-Use Micro-data: Overview and Research Problems," in (J. Domingo-Ferrer and V. Torra, eds.) Privacy in Statistical Databases 2004, New York: Springer, 231-247, http://www.census.gov/srd/papers/pdf/rrs2004-06.pdf .

Winkler, W. E. (2004c), "Record Linkage: Overview of Recent Developments and Applications," in (S. Biffignandi, ed.) Combining Data from Different Sources – Applications of Record Linkage Methodology and Estimation Using Administrative Data, Rome: ISTAT, to appear.

Winkler, W. E. (2005a), "Overview of Record Linkage and Current Research Directions," U.S. Bureau of the Census, Statistical Research Division Report at http://www.census.gov/srd/www/byyear.html , to appear.

Winkler, W. E. (2005b), "Data Quality in Data Warehouses," in (J. Wang, ed.) Encyclopedia of Data Warehousing and Data Mining, to appear.

Winkler, W. E. (2005c), "Methods and Analyses for Determining Quality," to appear.

Winkler, W. E. (2005d), "Data Quality for Modeling, Analysis, and Data Mining," to be submitted.

Winkler, W. E. (2005e), "Methods and Analyses for Determining Quality," 2nd Keynote address at the 2005 *ACM SIGMOD Workshop on Information Quality in Information Systems* (available under Post Workshop Material at http://iqis.irisa.fr/).

Winkler, W. E. (2005f), "Modeling and Quality of Masked Microdata," *American Statistical Association, Proceedings of the Survey Research Section*, to appear.

Zayatz, L. (2005), "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update", Research Report Series (Statistics #2005-06), Statistical Research Division, U.S. Census Bureau, Washington, D.C.

---

[i] This paper was written as part of a session organized to honor Pat Doyle at the 2005 American Statistical Association meetings in Minneapolis. Many ideas are derived from discussions with John Abowd, Pat Doyle and Laura Zayatz. Thanks also to Nick Greenia for extensive discussions on data quality and harm issues, Bill Winkler for alerting me to additional data quality and confidentiality literature, Miriam Heller and Sang Kim for their ideas about the relationship between cyberinfrastructure and confidentiality, Nancy Lutz for her help in developing the model and Guy Almes, Fredrik Andersson, Matt Freedman, Cheryl Eavey, Nancy Gordon and Dan Weinberg for their suggestions. The views expressed here are the author's and do not necessarily represent those of the National Science Foundation.

[ii] Pat Doyle, Maintaining Data Access and Dissemination in Light of Changes in Technology. Draft document, August 9, 2002

[iii] Margo Anderson, SBE/CISE workshop, March 15-16, 2005

[iv] *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Pat Doyle, Julia Lane, Laura Zayatz and Jules Theeuwes, North Holland, 2001.

[v] President's Information Technology Advisory Committee

[vi] A good example is the 228 page document on the design and methodology of the Current Population Survey 2002)

[vii] I am grateful to Bill Winkler for providing me with the workshop information

[viii] Cyberinfrastructure is a term coined by NSF to describe new research environments which exploit the newly available computing tools to the highest

available level. These include computational engines (supercomputers, clusters, workstations – capability and capacity), mass storage (disk drives, tapes, …) and persistence networking (including optical, wireless), digital libraries/data bases sensors/effectors, software (operating systems, middleware, domain specific tools/platforms for building applications), and services (education, training, consulting, user assistance). See Atkins, Daniel E. et al. (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure.* Arlington, VA: NSF for more information. Available at http://www.nsf.gov/cise/sci/reports/atkins.pdf

[ix] http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=13553&org=CISE&from=fund

[x] SBE/CISE workshop, Match 15-16 2005, http://vis.sdsc.edu/sbe/About

[xi] Radio frequency identification, or RFID, is a generic term for technologies that use radio waves to automatically identify people or objects. There are several methods of identification, but the most common is to store a serial number that identifies a person or object, and perhaps other information, on a microchip that is attached to an antenna (the chip and the antenna together are called an RFID transponder or an RFID tag). The antenna enables the chip to transmit the identification information to a reader. The reader converts the radio waves reflected back from the RFID tag into digital information that can then be passed on to computers that can make use of it. Source: http://www.rfidjournal.com/article/articleview/207

[xii] For the full report, see http://www.gmb.org.uk/shared_asp_files/uploadedfiles/95420EED-6333-4746-9BC0-432145FDD379_RegionalDistributionCentres.doc

[xiii] Some states (including California and Texas) are "open record" states. For example, California birth records for 1905–1995 are available on the state Web site and include the person's full name, birth date, sex, mother's last name, and county of birth. California's "non-identifying births summary" database for 1996-1997 contains information on person's county of birth, birth date, sex, race/ethnicity; mother's birth date, race/ethnicity, and state of birth; and father's birth date and race/ethnicity. Kentucky has a database on the Web containing records for 1.1 million marriages (1973–2002). The database contains the name, age, race, residence, and number of prior marriages for both the groom and the bride, as well as the date and county of the marriage and the marriage certificate number.

[xiv] This line of reasoning is heavily influenced by discussions with Pat Doyle and John Abowd as well as the work of Mark Elliott 2001

[xv] For a summary of the workshop, see Lane (2003)

[xvi] Stephen Fienberg was the social science coordinator of this session; Shankar Shastry the computer science coordinator.

[xvii] See Project IDS http://www1.cs.columbia.edu/ids/index101503.html

[xviii] http://abilene.internet2.edu/observatory/

[xix] I am grateful to Carl Landwehr for making me aware of this.

[xx] SBE/CISE workshop, Match 15-16 2005, http://vis.sdsc.edu/sbe/About