New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers

John Abowd
Cornell University

Julia Lane[1]
The Urban Institute

[1] This work is partially supported by the National Science Foundation Grant SES-9978093 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging (R01-AG18854-01), and the Alfred P. Sloan Foundation. Much of the paper is drawn from joint work and discussions with John Haltiwanger and Martha Stinson. We thank Fredrik Andersson and Simon Woodcock for helpful comments. All errors are our own.

I. Introduction

Data, and data access, lie at the heart of social science research. Billions of taxpayer dollars are spent in supporting the collection and dissemination of federal, state and local data, billions of dollars are spent in data analysis, and this, in turn, both informs scientific understanding of core social science issues and guides decision in how to allocate billions of dollars in social programs. Although an entire analytical infrastructure depends on the dissemination of high quality data, statistical agencies which have gone to great expense to collect such data, then deliberately destroy data quality –- often in ad hoc fashion -- in order to protect respondent confidentiality. Indeed, many statistical agencies spend millions of dollars, with concomitant respondent burden, to collect microdata, only to suppress substantial numbers of the resulting tabular output, and create tables with unknown statistical properties.

It is now apparent that new challenges threaten the ability of national statistical institutes (NSI's) to release high quality public use data files (see Doyle et al, 2001). Technological advances in computer capacity and matching technology combined with the explosion of online access to federal, state and local administrative records mean that NSI's must either severely degrade the quality of public use datafiles or refuse to release them in order to protect respondent confidentiality (see Yancey et al., 2003, Domingo Ferrer and Torra, 2003 for excellent reviews of matching technology). This has very serious practical consequences.

The response to this threat by the statistical community has been to develop new technical and non-technical approaches that will protect confidentiality but that will also maintain the same quality of statistical analysis than was possible using old techniques (see, for example the work by Agrawal and Srikaut (2000), which exemplifies the work on privacy preserving data mining). The NSI community is also responding to the issue – the Conference of European Statisticians recently established a working group to recommend approaches to micro-data access.

One very promising technical approach has been to develop multiply-imputed synthetic micro-data (Rubin 1993). This has the advantage of completely protecting individual confidentiality, as well as providing users with access to data wherever they wish, but imposes substantial data producer costs and has been resisted by the user community because of data quality concerns. Another approach has been to develop restricted access sites, which permit researchers to work on-site with micro-data (Dunne, 2001). Yet a third approach has been to develop remote access procedures, which has the advantage of reducing researcher burden, but which involves substantial investments in hardware and software. In addition, there is likely to be considerable bureaucratic resistance to adopting innovative techniques and algorithms to protect data transmission (Blakemore, 2001) – and by the time that resistance has been overcome, the techniques may be obsolete.

This paper describes a proposal to combine all three approaches: namely, developing inference-valid synthetic microdata which can be accessed at a restricted

access site, together with access to the "gold standard" analytical data set through a Research Data Center network[2]. It also describes the promise of the development of other datasets - particularly multiple public use files that can be created from the same underlying data that can be targeted at different audiences.

II. Background

Fienberg (2003) summarized the technical goals of disclosure limitation techniques as follows: (i) inferences should be the same as if we had original complete data; (ii) researchers should have the ability to reverse disclosure protection mechanism, not for individual identification, but for inferences about parameters in statistical models; (iii) there should be sufficient variables to allow for proper multivariate analyses and (iv) researchers should not only have the ability to assess goodness of fit of models but also be provided with most summary information, such as residuals (to identify outliers). The core guiding principle should be to generate released data that are as close to the frontier as possible. These principles hold just as much for micro-data as for synthetic data.

Most of these principles are obeyed with synthetic datafiles (see Muralidhar and Sarathy (2002), and Abowd and Woodcock (2003) for reviews). While the approaches vary (one approach is to shuffle data; another is to develop samples composed of draws from the posterior predictive distribution of the confidential data, given some conventionally disclosure-controlled data), a major advantage is that the synthetic data contain exactly the same statistical information as the micro data, which satisfies Fienberg's first principle. In the second approach, while the synthetic data implicates (described below) are not identical, the analyst can use the between implicate variation to measure the extent to which confidentiality protection made the inferences less precise, which satisfies the second and fourth principles. The release of sufficient variables, principle three, is discussed below.

But the use of synthetic data as a substitute for public use files produced using conventional disclosure limitation techniques has not caught on with the user community. A major problem has been the concern that the results produced from synthetic data will not be the same as those from the "real" data. The only way to substantively address this is to compare the results from synthetic data products with the results on the "gold standard" confidential source file. This poses serious constraints for a number of reasons. First, access to the "gold standard" file is, by definition, highly restricted. Second, because there are typically many different possible uses of the micro-data files, even if analysis on the synthetic datafiles will be "close" to what is achieved using the "gold standard" files with one specification, researchers have reasonable concern about whether analysis be "close" using alternative specifications.[3]

---

[2] More detailed technical information is provided in a related paper Abowd and Lane (2004) – an early version of which was presented at Statistics Sweden, August 2003.

[3] Although this may be due to a lack of researcher familiarity with the disclosure limitation approaches currently in practice – and the degree to which increasing protection has affected data quality and inference reliability.

An obvious solution is to develop a two-part access protocol. The first part is to create a remote access site – a virtual Research Data Center (RDC) - which can provide access to the full metadata repository of information, together with the synthetic data. Researchers can use such a site to gain familiarity with the dataset structure, develop code, and estimate analytical models. Because the data are synthetic, the statistical institute supplying the data to the remote site has to invest considerably less in protection technologies, which should dispel some of the concerns raised by Blakemore (2001)

The second part is to then re-estimate the models the models can be re-estimated at an RDC on the "gold standard" file. The comparison of the two sets of estimates can be distributed as widely as possible – each analysis will provide an increment to the common body of knowledge as to what works and what doesn't. This approach is described in the following sections.

III. The New Approach

i) The value added of synthetic data approach

One attractive feature of the synthetic data approach is that it can be used to create multiple public use files can be created from the same underlying data - targeted at different audiences. For example, a demographic dataset such as the Survey of Income and Program Participation (discussed below) has at least two important user constituencies. One constituency is interested in modeling the participation in welfare programs that are state-specific, with state specific qualification criteria – in which case geography is critical. Another constituency is interested in modelling retirement decisions – in which case date of birth is critical. In another example, some users of business data (such as transportation agencies) are particularly interested in geographic detail, while others are interested in industry detail (such as industry analysts). Providing both levels of detail on the same data set immediately re-identifies important businesses. Yet jointly releasing both geography and date of birth or geography and industry creates serious disclosure risk, and hence statistical agencies typically reduce the quality of one or the other variable (or both) – reducing their utility to both sets of users. However, synthetic data could be used to produce two separate data sets that can not be re-linked for such re-identification.

Another attractive feature of synthetic data is the ability to assess the biases in the protection system and the potential to correct public use products – since prior releases of synthetic data do not compromise proposed new releases. This aspect can be facilitated by means of the development of a restricted access data center and access to the "gold standard" files at the national statistical institute headquarters.

There is, of course, some justifiable skepticism that synthesized data might hide important relations that a direct use of the confidential data would reveal. This is especially important if results are downwardly biased – since this may discourage further research. This makes the development of a feedback loop from the synthetic data to the

confidential microdata essential to develop confidence in these products and to ensure their continuous improvement – which is what is proposed here.

ii). The "Virtual" RDC

A sensible approach for facilitating high quality research is to maintain the data in a secure, restricted access environment, but widely distribute synthetic data through a restricted access remote site. Because the simulated data can be used at less secure sites than the statistical agency itself, researchers can develop an understanding of the structure of the datasets and use simulated data to develop code and estimate basic relationships before sending the code to the an official secure site to estimate the underlying relationships from the actual confidential data.

If multiple users can access the same dataset, and build on an existing database infrastructure, there are numerous advantages. Results can be replicated or expanded – which is a critical condition for scientific validity. Researchers can use existing datasets to cut the analysis in different ways, with different foci, which develops a broader understanding of the generalizability of results. In addition, the common use of similar dataset builds a common body of knowledge, as has been the case with public use files such as the Public Use MicroSample for the Decennial Census and the Current Population Survey. [4].

The cornerstone of the dissemination system is the virtual RDC, a replica of the research environment on the Census RDC network that uses synthetic data and the exact programming environment of the RDC network to permit researchers to develop research proposals and to interact with key Census employees. The virtual RDC can be used for primary research as confidence is built in the validity of the synthetic data for analysis of particular types of programs. More importantly, it can be used as an incubator for proposals to analyze the confidential data. Researchers can directly benefit from the fact that the structure of the synthetic data and the structure of the "gold standard" confidential data were identical. The researcher would develop the proposal in the same environment as a real RDC, thus guaranteeing that the tools needed to do the modeling were available and working properly.

A well used precursor to this model is the Cornell Restricted Access Data Center (CRADC, part of CISER at Cornell). The CRADC, which was developed under an NSF Social Data Infrastructure grant, as well as a National Institute on Aging grant, is the model for the virtual RDC. Authorized users access data from authorized providers using a "window" on the CRADC machines (which appear to be ordinary Windows computers to the user). The CRADC provides a complete research and reporting environment that fully supports collaboration among authorized users of the same data.[5] Although the CRADC is a reasonable model for a virtual RDC, the virtual RDC goes farther. Real RDCs operate with "thin client" interfaces to the RDC computing network, a specialized

[4] Indeed, a very powerful case for this approach has been made by Soete and ter Weel, 2003

5 Technically, all of the Census products on the CRADC are "public use" files; that is, they have been approved by the Census Disclosure Review Board for general distribution.

Linux environment. The virtual RDC will provide an exact replica of the supercluster computing system that we will implement to create the synthetic data and support the complex modeling on the "gold standard" and synthetic data.

The Census Bureau has already agreed to support an advisory panel of ten experts and users. Their role will be to provide regular (three times/year) feedback on the choice of data files to be synthesized and the quality of the data synthesizers.

iii) Research Data Centers

An important component of developing a new confidentiality protection system is to develop a research data center (RDC) network in which the quality of the new data product can be tested. The more sites that are available and accessible, the greater the ability of the scientific community to build the core common body of knowledge necessary for the acceptance and use of the new data product.

The existence of such a network is, of course, critical whether or not synthetic data approaches are adopted. An important consequence of the increasing threat of re-identification is that more and more noise is being added to public use datasets – with analytical consequences that would be unknown without access to the underlying confidential data. Since noise addition biases coefficients towards zero, researchers might, for example, incorrectly conclude that earnings differentials by race and sex had vanished over time – rather than realizing that more noise had been added over time!

The basic structure of the RDC network in the United States is well known, and described in both Dunne (2001) and on the Center for Economic Studies website (www.ces.census.gov). Briefly, RDC's enable external researchers to access micro-data under strict security protocols. All researchers must become Special Sworn Status employees of the Census Bureau (which involves fingerprinting, an FBI check, and an oath to protect the confidentiality of respondents – which, if broken, subjects the researcher to the penalty of a $250,000 fine and/or 5 years in jail). The researcher must document which files will be accessed, which variables used, and for which period of time. The researcher must also demonstrate that the predominant purpose of the research is to improve Census Bureau censuses, surveys and inter-censal population estimates, and provide a post-project certification that this has been achieved (see Greenia, 2004).

IV. Application

The LEHD Program, in conjunction with an interagency committee that includes the Social Security Administration, Internal Revenue Service, Congressional Budget Office and other parts of the Census Bureau, is developing a public use file containing data from the Survey of Program Participation (1990-1996 panels) and Social Security administrative/tax data (W-2 information separately by employer, Summary Earnings Records, Master Beneficiary Records, Supplemental Security Records and Form 831 Disability Records. The confidentiality protection of this public use file is particularly challenging because the SIPP source records cannot be re-identifiable in the existing SIPP public use files; that is, this new public use file must be used independently from

the existing SIPP public use files.

The development of the SIPP-SSA public use file has provided much needed experience in developing the layers of the confidentiality program. Since this public use file is targeted at retirement and disability research for national programs, all geography has been removed from the public use portion. Of course, the geography is still present on the internal files, so RDC access can be provided for those variables. Removal of the geography was necessary to limit the potential for re-identifying SIPP source records in the existing SIPP public use files. Preserving marital relations as well as basic demography and education variables provided the maximum extent to which conventional identity disclosure control methods could be used. The interagency committee thought that linking a handful of extremely coarse demographic and educational variables from the SIPP to the massive amounts of administrative and tax data was not the most effective method of providing access to these data.

As an alternative, a layered approach was adopted. Successive, confidential versions of the linked data including a long list of proposed variables from the SIPP and all of the administrative variables from SSA (including the tax data) were developed. Researchers at Census, SSA, IRS, and CBO are studying the variables in these files, deemed "gold standard" files because they contain all of the original confidential data. Once the research teams are satisfied that the gold standard files adequately provide for the study of statistical models relating the variables of interest from the SIPP and the administrative data, a variety of potential public use files will be produced using the methods described in this section of the proposal. The same research teams will then assess the bias and loss of precision from the various techniques. Other research teams will assess the identity and attribute disclosure risks from each of the methods. The committee will then be equipped with reasonable quantitative measures of the disclosure risks, scientific biases, and losses of precision associated with feasible implementations of these new confidentiality protection techniques. It is expected that a public use product will be available within two years. Interim products include full RDC support for the gold standard files, which contain links that permit the RDC use of any variable in the existing public use SIPPs. The SIPP-SSA public use file is not a static product. We fully expect the interaction of RDC-based researchers with the data to provide much needed feedback to the process of variable selection and confidentiality protection for such files.

Summary

The continued distribution of public use data-files is clearly threatened by the increased re-identification risk associated with both technological advances in linking software and the widespread availability of administrative records. It is clear that new approaches to developing public use data files must be investigated. This paper suggests the adoption of a three-tiered approach that combines both technical and non-technical approaches. The technical approach – the creation of synthetic datasets – could, in principle, permit the creation of multiple public use datasets from a single underlying confidential file that could be customized for multiple different constituencies. The non-technical approach is to combine the use of an already well accepted RDC network with that of a "Virtual"

RDC to both reduce the access costs and develop a common body of knowledge about the quality of the results generated from the analysis of synthetic data files relative to that from confidential micro-data. While the initial results have been quite promising, more extensive research is ongoing.

References

R. Agrawal, R. Srikant, `Privacy preserving data mining', Proc. 2000 ACM SIGMOD Int'l Conf. Management of Data, ACM Press, 2000, pp.439-450.

Abowd, John M. and Simon Woodcock, "Disclosure Limitation in Longitudinal Linked Data," in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), (Amsterdam: North Holland, 2001), 215-277.

Abowd, John M and Simon Woodcock, "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data", mimeo, Cornell University, 2003

Blakemore, Michael The Potential and Perils of Remote Access, in P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. M. Zayatz, Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies, Elsevier, 315-340.

Domingo-Ferrer, J and V Torra "Advanced Record Linkage for Disclosure Risk Assessment" mimo, presented at National Science Foundation May 21-22, 2003.

P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. M. Zayatz, Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies, Elsevier, 315-340.

Dunne, Timothy, "Issues in the establishment and management of secure research sites" in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), (Amsterdam: North Holland, 2001).

Fienberg, Stephen, "Allowing Access to Confidential Data: Some Recent Experiences and Statistical Approaches" presented at Statistics Sweden, August 21 2003

Greenia, Nicholas "Developing Adoptable Disclosure Protection Techniques: Lessons Learned from a U.S. Experience" in this volume, 2004.

Muralidhar, Krishnamurthy and Rathindra Sarathy, "Application of the Two-step Data Shuffle to the 1993 AHS Data: A Report on the Feasibility of Applying Data Shuffling for Microdata Release," research report prepared for the Census Bureau (June 2002).

D. B. Rubin, Discussion on statistical disclosure limitation, Journal of Official Statistics, vol. 9, no. 2, pp. 461-468, 1993.

Soete, Luc and Bas ter Weel, "ICT and Access to Research Data: An Economic Review", Maastricty Economic Research Institute on Innovation and Technology, mimeo, June 2003

Yancey, William E., William E. Winkler and Robert H. Creecy, "Disclosure Risk Assessment in Perturbative Microdata Protection," Research Report Series Statistics 2002-01, available online at http://www.census.gov/srd/papers/pdf/rrs2002-01.pdf, cited June 11, 2003.