Journal of Official Statistics
Vol. 9. No. 2, 1993, pp. 503–506

Discussion
Computer Security

Gerald Gates[1]

The paper by Sallie Keller-McNulty and Elizabeth Unger presents a new approach to disclosure limitation – at least as far as statistical research data are concerned. The authors' approach is to apply data protection techniques developed for the computer science profession to handle similar problems facing statisticians. In short, the methods deal with inference security in large, static databases by applying controls on the system rather than the data.

When the U.S. Census Bureau releases its public-use microdata files, it applies disclosure limitation measures (controls) to the data. These techniques, such as rounding, grouping into categories, topcoding, adding random error, or data suppression, are designed to virtually eliminate the possibility that an intruder could identify a respondent based on unusual characteristics or similar information available from other sources. Controlling the data has proved to be the most flexible means of providing research data to many users with different resources and interests. There are no restrictions on who may purchase a public use microdata file or where, or how, these files may be used. In addition, they do not need to be secured since they are no longer considered confidential.

Although there will always be an interest in public use microdata files, there is an increasing number of occasions where these products are not suitable because of the difficulty in applying adequate controls while maintaining the desired level of data utility. For instance, the merging of administrative records information with survey data, the development of powerful computer matching programs, and the existence of large private and public databases with identical variables are critical factors used in deciding what data to release. Computer matching has, at the same time, become both a useful tool to statisticians and a threat to data security. In response to this "Catch 22," survey organizations have initiated many administrative solutions (see Jabine 1993) and explored some legal options (Gates 1988).

Keller-McNulty and Unger offer a very different solution. Instead of controlling the data, why not control the operating system and database management system which access the data? This has been the basis for computer inference security applied to most dynamic databases and could have applicability to static statistical databases. I can think of two experimental methods that have taken this approach. The Luxembourg Income Study (LIS)

[1] U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.
* The views expressed here are attributable to the author and do not necessarily reflect those of the U.S. Census Bureau.

allows users to access restricted databases on-line in a controlled environment (Cigrang and Rainwater 1990). Users are issued passwords and must interact with a gatekeeper (both machine and human) that controls the output. A second approach accomplishes the same thing but attaches the system controls to the data. An example is the National Center for Education Statistics (NCES) recent release of encrypted data from the National Survey of Postsecondary Faculty on CD-ROM (Wright and Ahmed 1990). The system software on the disk restricts the output to weighted counts of cross-classified information. The major difference between the on-line and media-based methods is the amount of control retained by the survey organization.

Using database systems to provide inference security allows the survey organization to restrict the output data rather than the input data. This has the distinct advantage of maintaining the covariance and other analytical structures of the data so that various analyses can be done. The disadvantages lie in unforeseen system vulnerabilities that allow confidential data to be obtained or deduced. (This is not unlike the risks associated with failure to apply the appropriate disclosure avoidance protections to public use microdata.) In addition, these approaches have yet to explore thoroughly the potential risks of reidentification through sequential queries of the database, as noted by the authors.

The Census Bureau has no experience with this method of disclosure limitation. Primarily, there has been little incentive to pursue it since the microdata products released by the Census Bureau are ASCII files, not in a database format. Recent developments involving the Survey of Income and Program Participation (SIPP), however, may encourage us to do some research in this area.

Recently, the University of Wisconsin developed, with NSF funding, the "SIPP On-Line Database." The goal of the project, which was recently transferred from Wisconsin to the Census Bureau, was to provide on-line access to SIPP public use microdata in database form. Access to the data was provided by the standard database language SQL. The advantage to this format is that it allows users to select the small set of data they need from up to 25 reels containing an entire panel of SIPP interviews.

The SIPP On-Line Database consists of the same information approved for release on the SIPP public use microdata files (i.e., the protections are applied to the data). In addition, for security purposes, the computer on which these data reside is not connected to any other Census Bureau computer. Having a user file, in database form, on-line, presents us with a unique opportunity to develop a system to control user access and limit disclosure.

Recently, the U.S. Census Bureau discontinued providing the database program with the SIPP On-Line Data Base. We found that the database program was too slow and too limiting for most users. Consequently, users extract data to their own system for use with SAS or other software. Needless to say, this has limited possible research involving database security techniques such as those proposed by Keller-McNulty and Unger.

Despite this temporary setback, I believe that these techniques offer promise for expanding access to research microdata and I commend the research of Keller-McNulty and Unger. For our part, we have explored many ways to providing "safe" access to our statistical research microdata and we will need to consider computer inference security as part of our

overall data release strategy. In addition, the experiences of the LIS and NCES will give us a starting point for further research.

I would like to conclude by offering some observations about our current data release practices and what the future might hold.

- The situation today is not the same as it was ten years ago when the Census Bureau initiated its formal microdata review process. Computers are smaller, faster, interactive, and connected. Computer matching software is more sophisticated and matching databases more plentiful. To keep pace, data protections are limiting the utility of microdata files. Other options must be considered.
- The current microdata review process cannot adequately account for the decision process of a potential intruder. Therefore, the release decision rests primarily on the risks associated with what is possible. This is most likely over-protective, but reasonable given the consequences of a wrong decision. We need to evaluate a more systematic approach to incorporating threat, as well as risk, into the microdata review process.
- Keller-McNulty and Unger suggest an interdisciplinary approach to the disclosure avoidance issue which makes intuitive sense. Computer science offers data analysts the opportunity to manipulate data quickly and cheaply. It can also offer data providers a powerful means to protect it.
- The costs for computer security measures will play an important role in choosing the best way to release research data. If significant human resources are required to implement a system-based inferential security program, the costs may exceed the benefit.
- Privacy issues, both real and perceived, must not be overlooked. Despite our assurances that on-line access is safe, any breaches which occur in any on-line system will have consequences for all such systems.

A final point I need to make relates to a misunderstanding about the Census Bureau's plans for releasing public use microdata from the 1990 census. Often the 1990 census data swapping process is referred to as a technique used in the release of public use microdata. The data swapping is, in fact, not a microdata disclosure limitation procedure but rather a disclosure limitation procedure for tabulations produced from the census results. Public use microdata files created from the 1990 census will require the same data masking techniques applicable to all microdata files, even though the data may also have been swapped.

References

Cigrang M. and Rainwater, L. (1990). Balancing Data Access and Data Protection: The Luxembourg Income Study Experience. Proceedings of the American Statistical Association, Section on Statistical Computing, 1–4.

Gates, G.W. (1988). Census Bureau Microdata: Providing Useful Research Data While Protecting the Anonymity of Respondents. Proceedings of the American Statistical Association, Section on Social Statistics, 235–240.

Jabine, T.B. (1993). Procedures for Restricted Data Access. Journal of Official Statistics vol. 9, (this issue).

Wright, D. and Ahmed, S. (1990). Implementing NCES' New Confidentiality Protections. Proceedings of the American Statistical Association, Section on Survey Research Methods, 446–449.