The Confidentiality Beasties: A Fable About the Elephant, the Duck, and the Pig

(Loosely adapted without permission from The Fox, the Chicken and the Bag of Grain by Aesop)

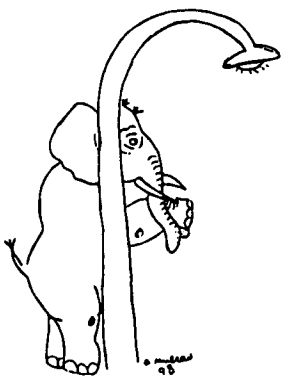
Fritz Scheuren and Jeri Mulrow, Ernst and Young, LLP

The problem of releasing public-use data while protecting the confidentiality and privacy of the individuals providing them is one of the most challenging facing statistical agencies at this time. The technology to collect and assimilate large amounts of data is now available via the Internet. Increasing sophistication continues in the field of computerized record linkage (Winkler, 1998); the methodology to protect and secure this information has not kept pace with the technology.

This paper is meant to capture the essence of what was said at the 1998 Joint Statistical Meetings in Dallas on the release of public-use data. The paper is divided into six sections: introduction, problems, fixes, alternatives, nexts, and an acknowledgment These correspond roughly to the slides used at the Dallas meetings. Also introduced are three confidentiality "beasties": the elephant, the duck, and the pig. These confidentiality "beasties" wander in and out of our remarks.

■ Problems

This is more appropriately titled the Elephant. The crux of the problem is how to hide an elephant behind a lamppost without distorting the dimensions of the elephant or the lamppost. This is not done easily either visually or mathematically/statistically.



Obvious identifiers are removed from public-use files (PUF), such as name, identification or Social Security number, and geography in the form of address and phone number. Oftentimes, variables are top-coded or grouped, such as income and race. As variables are removed or distorted to create PUF's, the probability of identification goes to zero. However, analytic utility also diminishes. Users of PUF data want the whole elephant, not a slimmed down, colorless elephant without a trunk, ears, tusks, or loud trumpeting voice. They do not want a lamppost as big as a brick wall to hide the elephant either. In other words, cumbersome access arrangements need to be avoided.

Longitudinal and heavy-tailed data present further complications to the release of public-use data. As more knowledge is gained over time about a particular record in a file, the more information is available for identification purposes. With very few exceptions, heavy-tailed data, such as business data, are not released for public use. Spruill (1982, 1983) began some interesting work in this area, but it did not result in a PUF. Much more satisfying solutions are still needed in these areas.

The need for variance calculations creates a problem. In many large surveys, geography is linked with the sample design and the weight associated with each record. Providing the proper information to correctly calculate variances can provide the information needed to pinpoint geography, which can provide the information to make an identification (Hinkins and Scheuren, forthcoming).

Gates (1998) provides a quote by Sally Katzen, the administrator of the U.S. Office of Management and Budget's Office of Information and Regulatory Affairs, at a 1997 press conference that sums up the ease of access to information:

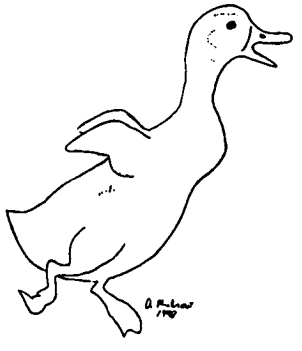
> "Today, a high school student in Pittsburgh, Pennsylvania, has better access to Federal statistics than a top Government official five years ago."

Data stewards have a new and harder challenge. The expansion beyond a small group of researchers, for example, known by name and known to be trustworthy, fuels one concern. Independently, a growing distrust in government is arising, and this fuels another.

In the current climate, data producers are getting more and more nervous. Data miners see more potential commercial uses. Data providers are less willing and more aware. But enough of problems. Hope you liked the picture anyway, if not the problems.

■ Fixes

This may be more appropriately titled the Duck. If it looks like a duck, swims like a duck, quacks like a duck, then it is most likely a duck. That is, some of the 'fixes' may not be fixes, and the data may still contain enough information to lead to an identification.



The Checklist on Disclosure Potential of Proposed Data Releases provides an excellent and systematic place to rebuild or maintain access. We can credit Ginny deWolf, Laura Zayatz, and Al Zarate (among others) with this fine effort. It should be the starting point for creating any public-use file.

Another good, but short-term fix to preserving confidentiality while allowing for complete analytic utility is the creation of research data centers. Both the Census Bureau and the National Centers for Health Statistics have taken this step.

The use of signed agreements, as advocated by some, is a step in the right direction, but not a fix in itself, since they are not enforceable. Once the data leave an agency, it becomes nearly impossible for the agency to completely monitor potentially inappropriate data use, such as identification. Thus, unscrupulous persons may decide to sign and use the data for commercial purposes without much fear of being caught.

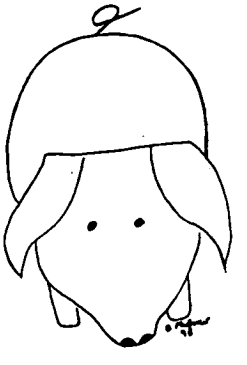
There has been some good theoretical work in the area of masking techniques. For instance, Kim (1986) and McCuckin and Nguyen (1988) have studied the addition of noise to data. Paass (1988) identifies different types of intruders and conditional types of protections against these type of intruders. Lambert and Duncan (1986, 1989) have contributed significantly.

Some agencies have staged break-ins to identify weaknesses in their public-use files. (We used to do this at IRS.) At first, agencies often use internal people to break in; then, they move on to contractors who can provide a different viewpoint (which is so important). Although the specifics cannot be spelled out, it is sufficient to say that every break-in that we have been privy to (at IRS and elsewhere) has been successful in one manner or another and has yielded surprisingly simple areas for improvement.

Lastly, other less pessimistic views should be mentioned. In particular, the National Academy Report *Private Lives and Public Policies* (1993) is optimistic on some points (but see also Scheuren, 1995).

■ Alternatives

This may be more appropriately titled the Pig. Should the goal be to create a pig without the squeal? If science could create such a creature, then would it be real? What would it taste like? If data producers could provide disclosure-risk-free data, would researchers have the analytic utility they desire?



There are several roads to choose from. One is to continue to improve the format of public releases. We also could certainly take the path of building an electronic query system. The Data Access and Dissemination System, or DADS, at the Census Bureau is an example. Steel and Zayatz (1998) make a compelling case for this approach. However, more study appears to be needed. Goes, Gopal, and Garfinkle (1998) take a slightly different path by using camouflage techniques.

Kennickell (1998) has taken still another road. He is among those truly working on creating the pig without the squeal. Multiple imputation (Rubin, 1987) is his tool of choice here. Synthetic microdata via multiple imputation techniques are not in their infancy, but they are certainly not fully grown up either. Clearly, more work is needed in this area, including giving researchers an opportunity to decide 'how it tastes.'

We are still early in the discovery process--too early to make predictions about outcomes. Who knows? After all, look what has been done with lambs. Whatever happens, we could be in an exciting time.

■ Nexts

We have touched on a few alternative approaches. Now, where do we go from here? In the world of continuous improvement, there is always a next step. Sometimes, there are multiple next steps depending on the situation—as is the case here.

We have not yet touched on the notion of confidentiality editing, but believe it is a good idea. It is consistent with Dalenius's data-swapping idea. As Steel and Zayatz (1998) noted, editing takes place on the file before production of published totals and release of microdata, thus avoiding the paradoxes that have existed between these two in the past. It is certainly worth pursuing.

As was stated earlier, the use of the Interagency Confidentiality and Data Access Group Checklist is a great teaching tool. It can still be improved, though. The process it outlines is not fully auditible yet, for example.

In keeping with the idea of a self-auditing system, we suggest that agencies consider obtaining an external process certification of their disclosure avoidance methods ('a la ISO9000.) The certification process can provide an agency assurance that a reliable approach was followed in releasing the public-use data and that care was taken to avoid known disclosure problems. It also affords an agency one last look, from the outside by a friendly agent, for potential chinks in their armor. The key word here, by the way, is external.

The use of data centers as a way to allow immediate access to sensitive data, such as business or linked data, makes sense. The Census and NCHS are taking the lead here. An alternative to researchers who cannot travel to a data center is the idea of licensing. NCES has done this. It should be approached cautiously, especially if there is no legislative structure of the sort NCES has. Again, once data leave the grounds, it is impossible to know how they are being used. Whatever is done, it is critically important to inculcate into researchers the "culture of confidentiality" that exists in statistical agencies.

We must continue to experiment with different methods. Computing and data access conditions have changed rapidly in the past few years. There is no reason to think that the future will not change just as rapidly if not faster. But it can be fun to anticipate the unanticipated.

Record linkage research is essential. Data mining is growing fast and is adding to the pressure on often outdated confidentiality protections. Measuring this risk is key. To be specific, consider the possible experiment using record linkage techniques:

(1) There should be a multiagency attempt to break in to existing files.

(2) The usual suspects, I mean agencies, should participate, but not be explicitly identified.

(3) The break-in would be done with existing public-use data, using modern linkage methods.

(4) Naturally, the intruder, a friendly one under con-

tract, perhaps E&Y, would have to operate legally, even while conducting unethical behavior.

(5) The list of possible identifications along with the specific method of attack would be provided back to each participating agency.

(6) Each agency alone would know which identifications were correct and which were not.

(7) A pooled result would be provided publicly, perhaps even by a third independent group, such as the Federal Committee on Statistical Methodology.

(8) Under this scenario, the cost of quantifying the risk would be shared, and the perception problem mitigated.

We have not discussed privacy concerns or physical security in this article. Research needs to continue in these other two areas alongside the confidentiality research proposed here.

Finally, to close, I would like to leave you with words from Senator Moynihan's speech earlier in these meetings:

> "You never do much about an issue
> until you put a number on it."

We need to move from implicit to explicit assumptions; from qualitative processes, such as checklists, to quantitative ones, such as the experiment that has been suggested. Thank you.

■ An Acknowledgment

We would like to acknowledge Alexandra Kay Mulrow (age 11) for her contribution of the drawings for this paper.

■ References

Cooper, J.; Merrell, D.; Nucci, A.; and Resnek, A. (1998), Protecting Confidential Data at Restricted-Access Sites: Lessons Learned from Census Bureau Research Data Centers, presented at 1998 Joint Statistical Meetings in Dallas, TX.

Duncan, G. T. and Fienberg, S. E. (1997), Obtaining Information While Preserving Privacy: A Markov Perturbation Method, *1997 Proceedings of the American Statistical Association, Survey Research Methods.*

Duncan, G. T.; Jabine, T. B.; and deWolf, V. A. (1993), *Private Lives and Public Policies*, National Academy Press, Washington DC.

Duncan, G. T. and Lambert, D. (1986), Disclosure-Limited Data Dissemination, *Journal of the American Statistical Assocation.*

Duncan, G. T. and Lambert, D. (1989), The Risk of Disclosure for Microdata, *Journal of Business and Economic Statistics.*

Gates, G. W. (1997), Information Privacy: Redefining the legal and ethical framework when the physical boundaries disappear, *1997 Proceedings of the American Statistical Association, Government and Social Science Statistics Sections.*

Goes, P.; Goal, R.; and Garfinkel, R. (1998), Confidentiality Via Camouflage: The CVC Approach to Database Security, presented at 1998 Joint Statistical Meetings in Dallas, TX.

Hinkins, S. and Scheuren, F. (forthcoming), An Inverse Sampling Algorithm for Confidentiality Protection.

Kennickell, A. (1998), Multiple Imputation in the Survey of Consumer Finances: Experience from the 1989-1998 Surveys, presented at 1998 Joint Statistical Meetings in Dallas, TX.

Kim, J. (1986), A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation, *1986 Proceedings of the American Statistical Association, Survey Research Methods.*

Martin, M. H. (1997), Monitoring Income for Social and Economic Development, presented at Evaluating Comprehensive State Welfare Reforms: A Confer-

ence at the Institute for Research Poverty at the University of Wisconsin-Madison, November 1996.

David, M. H. (1998), Killing with Kindness: The Attack on Public-Use Data, presented at 1998 Joint Statistical Meetings in Dallas, TX.

Paass, G. (1988), Disclosure Risk and Disclosure Avoidance for Microdata, *Journal of Business and Economic Statistics.*

Rasinski, K.A.; Timberlake, J.; Lee, L.; Porras, J.; and Mulrow, J. (1997), Producing a Public Use File: A Case Study, *1997 Proceedings of the American Statistical Association, Survey Research Methods.*

Rubin, D. (1987), Multiple Imputation, Wiley, New York.

Scheuren, F. (1995), A Review of Private Lives and Public Policy, *Journal of the American Statistical Association.*

Spruill, N. L. (1982), Measures of Confidentiality, *Statistics of Income and Related Administrative Record Research*, Internal Revenue Service.

Spruill, N. L. (1983), Confidentiality and Analytic Usefulness of Masked Business Microdata, the Public Research Institute, Alexandria, VA.

Steel, P. and Zayatz, L. (1998), Disclosure Limitation for the 2000 Census of Housing and Population, presented at 1998 Joint Statistical Meetings in Dallas, TX.

Winkler, W. E. (1997), Producing Public-Use Microdata That Are Analytically Valid and Confidential, presented at 1997 Joint Statistical Meetings in Anaheim, CA.

Winkler, W. E. (1998), Record Linkage Methods for Administrative Lists, presented at 1998 Joint Statistical Meetings in Dallas, TX.