Science, Confidentiality, *and the* Public Interest

John M. Abowd and Lars Vilhuber

O Privacy, Where Art Thou?

Aleksandra Slavkovic,
Column Editor

In the spring issue of *CHANCE*, Stephen Fienberg provided background and guidance on the interaction between privacy, confidentiality, disclosure, and harm. In this month's column, we will continue down that path, describing in more detail the benefits of providing data to public agencies and how public agencies navigate the narrow path between too much information disclosure on one hand and the release of useful information on the other.

Why Are the Services of Statistical Agencies Public Goods?

As Americans did last year, Canadians this year filled out their Census forms, although often they did so using their computers, rather than pencils. And Germans are being surveyed this year as well, although some of them may not need to touch either a keyboard or a pencil (more on these different methods of collecting data on people in a later column). So, in between updating their social-media wall and filling out a CV on a job-search website, why would people want to provide this information to the statistical agencies?

Statistical agencies in all countries exist to provide data to the general public and politicians. Recent statistics that have made the headlines include the following:

> "Hiring in U.S. Slowed in May with 54,000 Jobs Added" (*The New York Times*, June 3, 2011) – statistics provided by the Bureau of Labor Statistics
>
> "EHEC Epidemic in Germany: Where Are the Victims, and How Many" (*Stern*, June 15, 2011) – statistics provided by state agencies in Germany
>
> "Portugal Social Democrats Set to Win Election-Exit Polls" (*Reuters*, June 5, 2011, 3:02 p.m.)

To generate such statistics, firms (in the first example), hospitals (for the second example), and recent voters (for the third example) had to answer questions. A commonality of all the examples, though, is that the information benefits a large number of people. In fact, all can "consume" this information, without decreasing the utility of the information for others. And once the information has been released, it is difficult to exclude some uses. Economists call this a "public good."

Of course, if some people can obtain the information earlier than others, they may draw an advantage: selling shares on Wall Street earlier to avoid losses as the stock market drops as a consequence of the bad labor market news, as in the first example, or by simply being the first social-democrat to open a celebratory bottle of Champagne, in the third example. There may even be a profit motive behind the collection of the data, as is the case for exit polls, which are not typically run by government statistical agencies. But in all cases, the resulting information, once published, is useful to all people (citizens and non-citizens), allowing them to make informed plans and undertake informed actions.

How Do Public Goods Influence What a Statistical Agency Does with the Confidential Data It Collects as Part of Its Ongoing Operations?

Pure statistical agencies such as the U.S. Census Bureau and the Institut für Arbeitsmarkt- und Berufsforschung (IAB) in Germany don't enforce laws. Instead, they collect data from censuses, surveys, and administrative records. Usually, there is a specific statistical purpose the agency originally designed its data acquisition to address. For example, the data from the U.S. Census of Population are used for reallocating the seats in the House of Representatives and for redrawing the congressional districts. IAB, much like the Bureau of Labor Statistics in the United States, collects data from households and businesses to facilitate research on the labor market, which it places in the public domain.

Statistical agencies usually operate under a strict statutory mandate or voluntary pledge to protect the confidentiality of the respondents' data. This is where the tension of the public good problem rears its head. A pure statistical agency has no reason to exist if it does not publish data, and it should publish as much information as possible from the data it collects to maximize the "information re-use" benefits of the public good. But it cannot simply release the raw data because of the confidentiality considerations. So, statistical agencies do something to make the statistics they publish less informative about the entities that provided the data. This process is called statistical disclosure limitation (SDL). (See FCSM Discussion Paper 22, revised 2005, for an overview of statistical disclosure limitation practices in the United States.)

As their name suggests, disclosure limitation methods deliberately make the data an agency publishes in its tables, series, and research papers less informative than the tables, series, and papers one would publish if there were no confidentiality statutes or pledges. What the public users never see are the relationships, trends, and models that can only be discovered with unfettered access to the raw data. The stronger the disclosure limitation protocol, the less likely surprises emerge. These surprises are at the heart of the scientific discovery process. The potential to discover and publish some of these unexpected relationships is the primary source of researcher demand for access to the underlying confidential data a statistical agency collects.

How Does Research Benefit the Public Interest and the Agency?

The research centers inside statistical agencies, and the external researchers who use the confidential versions of the data, also publish public data—namely, the results they place in the scientific literature. Just one statistical agency, the U.S. Census Bureau, now releases hundreds of such papers each year, as reported in the *Annual Report of the Center for Economic Studies*. And many others in countries around the world do likewise.

Important science has come from this access to confidential statistical agency data. The earliest quantification of the magnitudes of worker flows in and out of unemployment was based on confidential data from the Current Population Survey. Later, the confidential microdata from the Census of Manufactures were used to establish the patterns for hiring and separation relative to net employment changes. Our understanding of the role of job creation and destruction in the macroeconomics of recessions and booms originated from the use of confidential establishment-level American data, which showed enormous employer heterogeneity in the two rates right through most business cycles—successful firms kept creating jobs and failing firms destroyed them, reallocating workers to more efficient uses in the economy. Confidential French data showed the direct link between the two flows (three workers hired and two separated to create a job; two workers fired and one hired to destroy a job, on average). Confidential data from a host of countries have confirmed that the presence of long-term personal and employer differences in wage rates not related to measurable variables is not a consequence of incomplete models but, rather, an intrinsic feature of modern economies.

Access Modalities and Synthetic Data

To protect privacy, agencies thus lock the data away, and only release certain subsets for public use. Researchers, on the other hand, may need the confidential data to generate new results, new information that neither the agency nor previous researchers has thought about. Recent examples include the discovery that top-coding (a form of disclosure control in which actual data items above the "top-code" are replaced by that value) was leading to potentially biased inferences about the rise in U.S. income inequality and that the importance of job creation by small businesses was a statistical artifact—it is young businesses, which happen often to be small, that create the most jobs. Thus, to promote this kind of research, statistical agencies have developed methods to let researchers access the confidential data, without letting the fox leave the coop with the hens.

Physical access. Researchers are given physical access to the data at the agency, itself. Thus, the researcher travels to the agency and is sworn to lifetime secrecy about what he or she will be shown. Then, in facilities that are either owned and managed by the agency or at locations where access is controlled by the agency, but other infrastructure is managed by research facilities (research data centers), the researcher gets to analyze confidential data under a need-to-know protocol. Output that the custodial agency deems safe to release is produced and published at the end of the typical research cycle. The researcher is usually free to view all intermediate output on the iterative path to identifying the right modeling strategy and responding to peer review.

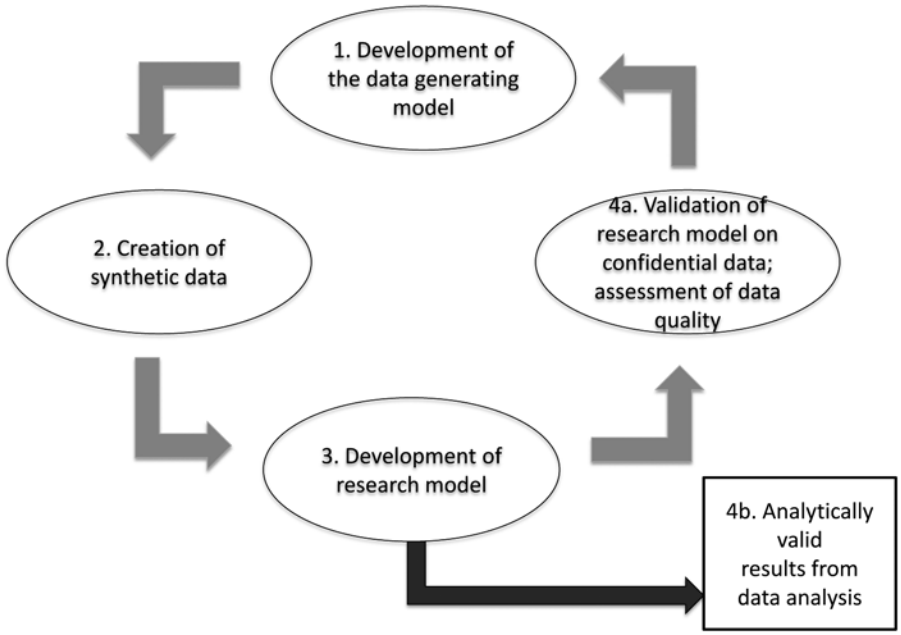

Figure 1.

The downside to this arrangement is that it is costly both for the agency, which needs to maintain such special-purpose facilities, and for the researcher, who typically has to travel to the relevant site and work with restricted computing facilities. Some of the inconvenience can be mitigated by creating satellite offices in multiple locations, reducing (but not eliminating) the travel costs for many researchers, but at an additional infrastructure cost.

Remote access by proxy. To address the travel problem, agencies have provided staff and facilities that will perform some of the tasks for the researcher. For instance, they may provide programmer services to researchers, such that the researcher only needs to provide instructions to the programmer and is then sent disclosure-limited output. This has the advantage of the researcher no longer needing to travel, but comes with both a monetary cost (to either the researcher or the agency for paying the intermediary) and a conundrum: To refine their model, researchers often want to see a lot of intermediate results, many of which would not be safe to release and are denied.

Remote access. In some cases, agencies have provided researchers with secure remote access. Researchers obtain an encryption device and sign legal promises to keep the data confidential. Using elaborate log-on procedures, researchers can then access the confidential data on their own schedule, from their own offices. This model is often used in the United States for data from surveys not run by statistical agencies, but is rarely used by North-American statistical agencies, in contrast to many European agencies that have started to make widespread use of such methods.

Recently, an additional method has gained traction:

Synthetic data. Data created by sampling from the posterior predictive distribution of the confidential data have been produced from several major confidential sources. The Survey of Income and Program Participation (SIPP) linked to lifetime earnings and benefits data from the Internal Revenue Service and Social Security Administration have been released in synthetic format. The Longitudinal Business Database (LBD), the establishment-level data developed from the Census's Employer Business Register and used to produce the Business Dynamics Statistics, also has been released as synthetic data. Additionally, synthetic data methods have been used in the Census Bureau's OnTheMap graphical mapping application.

For the SIPP and LBD data, the Census Bureau and Cornell University developed a unique access environment. The synthetic data are released as beta public-use products to a restricted-access computational server (Synthetic Data Server, SDS, at the Cornell Virtual Research Data Center). Access is granted to any person who applies proposing a project that uses only the data available in the synthetic versions. In exchange for developing their models on the synthetic data, the Census Bureau staff supporting these data performs the same statistical analysis on the confidential data using the programming developed on the SDS. The confidential data analysis is then subjected to the same statistical disclosure limitation procedures other model-based publications must use before it is released to the researcher.

The Scientific Feedback Loop and the Public Interest

Synthetic data servers are part of a general model of researcher feedback and data quality improvement that is illustrated in Figure 1. Starting with an existing confidential data set, agency staff

develops a data-generating model (1), which in turn allows them to produce one or more synthetic data sets (2). The quality of the generating model, and thus of the synthetic data, depends on the state of the science at the time of that first step. The synthetic data set is then released to the research community, which uses it to development a research model for a particular topic of interest (3). The research yields (hopefully) interesting results (4b), but at least in the initial phase, doubts persist as to the validity of those results, as they stem from a previously untested SDL procedure applied to a complex data set. This is where the validation of the results against the confidential data by the agency (4a) plays an important role: It generates confidence in the usefulness of the new data, if the results concord with those obtained from the synthetic data, or, if the results differ, allows the feedback loop to be closed by incorporating new information into the data-generating model (1). And the cycle starts again, leading to the release of improved synthetic data in the next round.

How Well Does This Work in Practice?

This model is precisely the one that was applied for the two synthetic data sets available at the Cornell Synthetic Data Server. The SIPP synthetic data have already undergone one complete cycle of this feedback process, and the newer version is much improved. The synthetic LBD data are still in the first cycle, but the first project has asked for replication of the results a scant three months after the data first became available. The types of questions asked of the data have been fairly broad. For the synthetic SIPP, the 20 projects since the first release of the data have investigated:

- The effect of Social Security benefit rules on the timing of divorce and wage gaps before and after divorce
- Trends in different measures of earnings volatility and instability; wage flexibility
- Associations between individual characteristics (including wages), uninsured spell lengths, and transitions into and out of health insurance
- Poverty of older women, using the full earnings and marital history
- Transitions over time between segments of the family income distribution and the family earnings distribution
- The male marriage premium
- Optimal extent and timing of pre- and post-tax savings, as well as the timing of Social Security distributions

However, the synthetic SIPP data also were analyzed for their data-generating features:

- An illustration of a data-sharing procedure that involves multiple parties (leveraging the fact that to put the data together required multiple agencies to share data)
- Multi-component hypothesis tests for use with partially synthetic data imputed in two stages (leveraging the fairly unique method of generating the synthetic data)

The five projects that have started to use the synthetic LBD since its release to the Cornell Synthetic Data Server in May 2011 have all focused on establishment and firm lifecycle issues. However, among the projects, some have had a stronger focus on using the synthetic data as a collaboration tool (among Census and non-Census researchers) and as a tool to prepare the analysis of the confidential data. Thus, while the use of the synthetic data for its statistical properties dominates, it is not the only use researchers have found for the data.

Final Thoughts

Although the benefits of scientific use of the confidential statistical agency data are now well understood, deep involvement of subject-area specialists in the testing and improvement of disclosure limitation methods is still rare. We think it should become a lot more common. Every use of a statistical product generates additional benefits from the public good. Participation of the research community in the improvement of the confidentiality protection just completes the link between science, confidentiality, and the public interest.

Further Reading

OMB-Federal Committee on Statistical Methodology Working Paper 22 (Revised 2005)

Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project I (November 2006), *www.census.gov/sipp/synth_data.html*.

CES-WP-11-04: Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database, *www.census.gov/ces/search.php?search_what=paps&detail_key=101943*

VirtualRDC, with additional information and links on synthetic SIPP, synthetic LBD, and the Synthetic Data Server, *www.vrdc.cornell.edu*