
What roles do remote servers and synthetic data play in the future of data dissemination?

New Approaches to Data Dissemination: A Glimpse into the Future (?)

Jerome P. Reiter

Many national statistical agencies and survey organizations disseminate microdata, i.e., data on individual units in public use data files. These data disseminators strive to release files that are safe from attacks by ill-intentioned data users seeking to learn respondents' identities or sensitive attributes, informative for a wide range of statistical analyses, and easy for users to analyze with standard statistical methods. Meeting all three goals is a challenging task. The proliferation of readily available databases, and advances in statistical and computing technologies, provide users with more and higher quality resources for linking records in released datasets to units in other databases. As a result, the risk of unintended or illegal disclosures is high and still rising. Microdata proliferation and statistical advances also enable and fuel the ambition of researchers. To address complex statistical questions, these users demand greater access to accurate data at fine levels of detail. Data disseminators thus find themselves in a difficult position: users pressure them to provide everything about the data, but disclosure risks pressure them to limit what is released.

Data disseminators that fail to prevent disclosures of individuals' identities or sensitive attributes can face serious consequences. They may be in violation of laws and therefore subject to legal actions; they may lose the trust of the public, so that respondents are less willing to par-



ticipate in their studies; or, they may end up collecting data of dubious quality, since respondents may not give accurate answers when they believe their privacy is threatened. As evident from the recently passed CIPSEA law in the U.S., these consequences are unlikely to diminish.

Given these trends, it is conceivable that, in the near future, data disseminators may not be willing or legally allowed to release any genuine public use microdata. Yet, the public's demand for high quality microdata is not likely to abate. Wide access to data facilitates advances in economics, public health, sociology, and many other areas of knowledge. Denying all public access to microdata would eliminate these societal benefits, which runs counter to the missions of most statistical agencies and organizations.

How can agencies and organizations continue to provide public access to microdata in a world where confidentiality constraints do not allow them to release genuine data? This article describes two data dissemination strategies for such a world, both of which are currently being researched by statisticians in academia and at national statistical agencies. The first is remote access computer servers, to which users submit requests for analyses and, in return, receive only the results of statistical analyses, such as estimated model parameters and standard errors. Confidentiality is protected, because the remote server never allows users to see the genuine data. The second is to release synthetic, possibly simulated, data that mimic the relationships in the real data. This approach has low disclosure risks since the released values are not the genuine data.

Discussion of these approaches is framed by two key questions. First, to what degree can these approaches protect data confidentiality? Second, how do these approaches affect the accuracy and types of analyses users can undertake? These two questions are relevant for any method of data dissemination, including disclosure limitation techniques used currently by many agencies and organizations. To provide context and motivation, we begin by examining some of these current approaches.

Current Approaches to Statistical Disclosure Limitation

Most data disseminators do the obvious things to protect confidentiality before releasing data, such as stripping unique identifiers like names, Social Security numbers, and addresses. However, these actions alone may not eliminate the risk of disclosures when key identifying variables—age, sex, race, and marital status, for instance—remain on the file. These keys can be used to match units in the released data to other databases. Most data disseminators therefore alter values of key identifiers, and possibly values of sensitive variables, before releasing the data. For example, they globally recode variables, such as releasing ages in five-year intervals or top-coding incomes, e.g. releasing incomes above \$100,000

as “\$100,000 or more”; they swap data values of keys for selected units—switching the sexes of some men and women in the data, for example — in hopes of discouraging users from matching, since matches may be based on incorrect keys; or, they add random noise to numerical data values to reduce the potential for exact matching on key variables or to blur the values of sensitive variables.

These strategies typically do not eliminate the risks of identification or attribute disclosures. Even when ages are collapsed in five-year categories, analysts may be able to identify records by examining rare combinations of other characteristics. With data swapping, typically most records are not altered to limit the harm to data utility. Unaltered records may be susceptible to disclosures. Similar risks apply when data are protected with added noise.

Applying these strategies adversely impacts the utility of the released data, making some analyses impossible and distorting the results of others. Analysts working with top-coded incomes cannot learn about the right tail of the income distribution from the released data. Analysts working with swapped sexes or races may obtain distorted estimates of relationships involving these variables. Analysts working with values that have added noise may obtain attenuated estimates of regression coefficients and other parameters. Accounting for these types of perturbations requires likelihood-based methods or measurement error models. These are difficult to use for nonstandard estimands and may require analysts to learn new statistical methods and specialized software programs.

As resources available to users continue to expand, the alterations needed to protect data with these techniques may become so extreme as to make the altered data practically useless. We next consider an approach that allows users to perform statistical analyses using unaltered data without releasing that data: remote access servers.

Remote Access Servers

Remote access servers are computers that house the collected microdata. Users submit requests for statistical output, but they are not allowed to see the data. When the request is deemed safe from disclosures, the server responds with parameter estimates, standard errors, and diagnostic measures of model fit. Several statistical agencies are developing or already use remote servers as part of their data dissemination strategies. Those agencies include the Australian Bureau of Statistics, Statistics Canada, Statistics Denmark, Statistics Netherlands, Statistics Sweden, U.S. Bureau of the Census, U.S. National Agricultural Statistics Service, U.S. National Center for Education Statistics, and U.S. National Center for Health Statistics (Rowland 2003).

Remote servers have advantages over releasing altered versions of the original data. First, analyses are based on the original data, and so are free from biases injected by data perturbation methods. Second, users of remote servers can fit standard statistical models; there is no need

to make corrections for measurement errors caused by data perturbations. Third, remote servers can protect confidentiality more effectively than releasing altered data, since no actual or close-to-actual values for individual units are purposefully released.

Although remote servers do not allow users to view the data, they are not immune to disclosure risks. Users may be able to submit models containing judicious transformations of variables that result in disclosures (Gomatam, et al. 2003). For example, suppose a user knows that a certain unit is in the dataset and possesses a unique value of some nonsensitive attribute, say $X=x$. To learn that unit's value of some sensitive attribute Y , the user could fit a regression using Y as the outcome and a single predictor variable that equals one when $X=x$ and equals zero otherwise. The resulting intercept and coefficient can be added to obtain the exact value of Y for that unit. Another attack is to submit models containing transformations of variables that create artificially extreme values for units with certain values of X . Such points pull fitted regression lines close to them, resulting in very accurate predictions of the

uals should not show any patterns when graphed against the values of the predicted values or the predictors themselves. Although very useful as tools for model diagnostics, the residuals, predicted values, and predictors cannot be released by remote servers. Otherwise, the user can obtain values of outcome variables by simply adding the residuals to the predicted values. A way around this problem for regression models was proposed by Reiter (2003a): remote servers can provide simulated diagnostics that mimic the patterns in real-data diagnostics. Users then can treat these simulated values like ordinary diagnostic quantities by examining scatter plots of simulated residuals versus simulated predicted values or versus simulated predictors, for example.

Many details of remote servers need to be ironed out before servers become a widespread method of data dissemination. User-friendly interfaces need to be developed for submitting models and reporting output. Capabilities for fitting sophisticated models need to be incorporated. Automated technologies for checking the disclosure risk of submitted models need to be created. This last task is

Although remote servers do not allow users to view the data, they are not immune to disclosure risks.

outcome for these units.

Disclosures can also occur from models that fit the data too well, even without unusual transformations. For example, suppose a particular, good-fitting regression for a sensitive outcome has a very small residual mean square error. The estimated coefficients can be used to obtain accurate predictions for units with known predictor values. Or, if all members with a certain pattern of predictors have identical outcomes, as may be the case for categorical outcomes, predictions from the fitted model will be exact for the units with that pattern.

To limit the risk of disclosure, servers can decline to provide output for models deemed too risky. The tricky part is deciding what models are too risky, and what models are legitimate inquiries from users. It is desirable that these decisions be made automatically by the server; performing manual checks of every proposed analysis can be time-consuming and expensive for data disseminators. Methods for performing such automated checks are currently being researched (e.g., Gomatam, et al. 2003).

The remote server should also provide some way for users to check the fit of their models. Unfortunately, releasing the usual diagnostic statistics can disclose values. For example, a common diagnostic tool in regression modeling is the residual, which is the difference between the actual value of the outcome and the value predicted by the fitted model. When the model fits well, the resid-

complicated by the fact that disclosures can arise from a series of seemingly innocuous queries, such as sequential queries that isolate certain units (Rowland 2003). Nonetheless, remote servers can be expected to play a central role in the future of data dissemination.

Synthetic Data

If data disseminators are not willing or not allowed to release genuine microdata, another approach is to release synthetic, or simulated, microdata that look like the genuine data. This was first proposed by Rubin (1993). To generate synthetic data, the agency or organization (1) randomly and independently samples units from the sampling frame to comprise each synthetic dataset, (2) imputes the unknown data values for units in the synthetic samples using models fit with the original survey data, and (3) releases multiple versions of these datasets to the public.

To illustrate how this might work in practice, let us suppose an agency has collected data on a random sample of 10,000 people. The data consist of each person's race, sex, education, and income. We assume the agency has a list containing all people in the population, including their race and sex. This list could be the one used when selecting the random sample of 10,000, or it could be manufactured from census tabulations of the race-sex joint distrib-

ution. We assume the agency knows the education and income only for the people who responded to the survey. In the first step to generating synthetic data, the agency randomly samples some number of people, say 20,000, from the population list. In the second step, the agency uses the collected data to estimate the joint distribution of income and education for each race-sex combination. The agency then generates values of education and income for the 20,000 synthetic people by randomly simulating values from these joint distributions. The result is one synthetic dataset. The agency repeats the process, say, 10 times, each time using different random samples of 20,000 people, to generate 10 synthetic datasets. These 10 datasets are then released to the public.

When the educations and incomes for the synthetic people are simulated from the true joint probability distributions, the synthetic data should have similar characteristics on average as the real data. There is an analogy here to random sampling. Some true joint distribution of education and income exists in the population. The observed data are just a random sample from that population distribution. If we generate synthetic data from that same distribution, we are essentially creating different random samples from the population. Hence, the user analyzing these synthetic samples is essentially analyzing alternative samples from the population.

The “on average” caveat is important: parameter estimates from any one simulated dataset are unlikely to equal exactly those from the observed data. The synthetic parameter estimates are subject to three sources of variation, namely sampling the collected data, sampling the synthetic units from the population, and generating values for those synthetic units. It is not possible to estimate the three sources of variation from only one released synthetic dataset. However, it is possible to do so from multiple synthetic datasets, which explains why multiple synthetic datasets are released. To account for the three sources of variability, the user estimates parameters and their variances in each of the synthetic datasets, and then combines these results using simple formulas described by Raghunathan, et al. (2003).

These methods adjust automatically for the size and number of synthetic datasets. The synthetic sample size need not equal the number of units in the collected data. Research shows that increasing synthetic sample size leads to relatively small gains in inferential accuracy pro-

vided the synthetic sample size is large to begin with. Increasing the number of released synthetic datasets can substantially improve accuracy.

How does releasing fully synthetic data prevent disclosures, and could confidentiality be compromised? Identification of units and their sensitive data from synthetic samples is nearly impossible. Almost all of the released, synthetic units are not in the original sample, having been randomly selected from the sampling frame, and their values of survey data are simulated. The synthetic records cannot be matched meaningfully to records in other datasets, such as administrative records, because the values of released survey variables are simulated

rather than actual. Releasing fully synthetic data is subject to attribute disclosure risk when the models used to simulate data are “too accurate.”

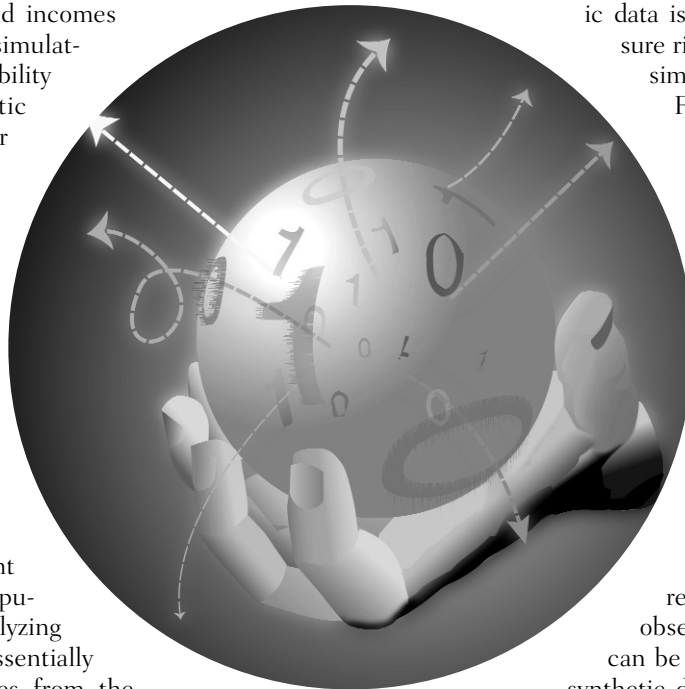
For example, when data are simulated from a regression model with a very small mean square error, analysts can estimate outcomes precisely using the model. Data disseminators can reduce this risk by using less precise models when necessary.

Synthetic datasets can have many positive data utility features.

When data are simulated from distributions that reflect the distributions of the observed data, valid inferences can be obtained from the multiple synthetic datasets for a wide range of estimands. These inferences can be deter-

mined by combining standard likelihood-based or survey-weighted estimates; the analyst need not learn new statistical methods or software programs. Synthetic datasets can be sampled by schemes other than the typically complex design used to collect the original data, so that analysts can ignore the design for inferences and instead perform analyses based on simple random samples. Additionally, the data generation models can incorporate adjustments for nonsampling errors and can borrow strength from other data sources, thereby resulting in inferences that can be even more accurate than those based on the original data. Finally, because all units are simulated, geographic identifiers can be included in the synthetic datasets, facilitating estimation for small areas.

There is a cost to these benefits: the validity of synthetic data inferences depends critically on the validity of the models used to generate the synthetic data. This is because the synthetic data reflect only those relationships included in the data generation models. When the models fail to reflect accurately certain relationships, analysts’



inferences also will not reflect those relationships. Similarly, incorrect distributional assumptions built into the models will be passed on to the users' analyses. This dependence is a potentially serious limitation to releasing fully synthetic data. Practically, it means that some analyses cannot be performed accurately, and that data disseminators need to release information that helps analysts decide whether or not the synthetic data are reliable for their analyses.

As of this writing, no agencies have adopted the fully synthetic approach, although several agencies have initiated research into the feasibility of this approach. High on the research agenda are investigations of semiparametric and nonparametric methods for generating synthetic data, and evaluations of synthetic data inferences on genuine datasets of varying structures.

Some agencies, however, have adopted a variant of the synthetic data approach called partially synthetic data (Reiter 2003b). Partially synthetic data include the units originally surveyed with some collected values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. The U.S. Federal Reserve Board protects data in the U.S. Survey of Consumer Finances by replacing monetary values at high disclosure risk with multiple imputations, releasing a mixture of these imputed values and the unreplaced, collected values. The U.S. Bureau of the Census protects data in longitudinal, linked datasets by replacing all values of some sensitive variables with multiple imputations and leaving other variables at their actual values. Partially synthetic approaches are appealing because they promise to maintain the primary benefits of fully synthetic data—protecting confidentiality while allowing users to make inferences without learning complicated statistical methods or software—with decreased sensitivity to the specification of the data generation models (Reiter 2003b).


The protection afforded by partially synthetic data depends on the nature of the synthesis. Replacing key identifiers with imputations makes it difficult for users to know the original values of those identifiers. Replacing values of sensitive variables makes it difficult for users to learn the exact values of those variables. Nonetheless, partially synthetic datasets are more susceptible to disclosure than fully synthetic ones. The originally sampled units remain in the released files, albeit with some values changed, leaving values that users can utilize for record linkages.

For fully or partially synthetic data to be accepted as methods of data dissemination, data disseminators will have to undertake a massive education campaign explaining to potential users the viability and limitations of the approaches. This campaign will succeed or fail based largely on evidence comparing analyses from synthetic and observed data. Developing accurate data synthesizers is a daunting challenge for statistical researchers, but it is worth pursuing. If agencies or organizations cannot release genuine microdata in the future because of confidentiality constraints, synthetic microdata may be one of the only

ways for researchers to get their hands on microdata.

The Future

The remote server and synthetic data approaches will not meet all analysts' statistical needs. Analysts seeking to use exploratory data analysis to search for complicated relationships may find remote servers too limited. Analysts seeking to fit models involving relationships not generated in the synthetic data—for example, high-order interactions involving complicated transformations of the data—will find the synthetic data inadequate for their modeling. Such analysts may have to apply for special access to the genuine microdata in restricted research data centers. These centers typically require analysts to sign special pledges of confidentiality, and all work using the data is done in the center. While restricted access data centers are undoubtedly part of the future of data dissemination, they are not a viable solution for wide access to public use data. Not all researchers live near centers or can afford to work at a distant center. These centers also are expensive to maintain, so that they are unlikely to proliferate.

Wide access to public use microdata has undeniable societal benefits that are worthwhile to maintain. Concerns over data confidentiality, which are growing as disclosure risks increase, threaten to extinguish those benefits. In the near future, agencies and organizations may not be willing or allowed to release genuine public use microdata. Statisticians in academia, government, and industry have recognized this coming problem and have proposed two potential solutions: remote access servers and synthetic datasets. Although the challenges to implementing these solutions successfully are great, the potential payoffs are even greater. Remote servers and synthetic data undoubtedly will play central roles in the future of data dissemination. 

References

- Gomatam, S., Karr, A.F., Reiter, J.P., and Sanil, A.P. 2003. Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. Technical Report, National Institute of Statistical Sciences.
- Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. 2003. "Multiple imputation for statistical disclosure limitation." *Journal of Official Statistics*, 19, 1-16.
- Reiter, J. P. 2003a. Model diagnostics for remote access regression servers. *Statistics and Computing* 13, 371-380.
- Reiter, J. P. 2003b. Inference for partially synthetic, public use microdatasets. *Survey Methodology*, 29, 181-188.
- Rowland, S. 2003. "An examination of monitored, remote microdata access systems." National Academy of Sciences Panel on Data Access workshop paper.
- Rubin, D.B. 1993. Statistical disclosure limitation. *Journal of Official Statistics*, 9, 461-468.