Data Confidentiality

Jerome P. Reiter
Duke University

Keywords
confidentiality, disclosure, microdata, imputation, synthetic

Abstract
When releasing data to the public, data disseminators typically are required to protect the confidentiality of survey respondents’ identities and attribute values. Removing direct identifiers such as names and addresses generally is not sufficient to eliminate disclosure risks, so that data must be altered before release to limit the risks of unintended disclosures. When intense data alteration is needed to ensure protection, the quality of the released data can be seriously degraded. This article reviews a disclosure limitation approach called synthetic data, in which values of confidential data are replaced with simulations from statistical models. Theoretical and empirical investigations have shown that synthetic data approaches have the potential to result in higher data quality than other disclosure limitation procedures, particularly when intense data alteration is necessary. The article discusses the main variants of synthetic data approaches, namely full synthesis and partial synthesis. It includes discussions of synthetic data generation and disclosure risk assessment.

Many national statistical agencies, survey organizations, and researchers disseminate microdata, i.e., data on individual units, to the public. Wide access to microdata facilitates advances in science and public policy, encourages replication of findings, enables students to train on genuine datasets, and helps citizens to stay informed about their society. Often, however, data disseminators cannot release microdata in their collected form, because doing so would reveal some respondents’ identities or values of sensitive attributes. Data disseminators therefore typically alter the collected data before release. Common strategies include recoding variables, such as releasing ages in five year intervals or geographies at high levels of aggregation; reporting exact values only above or below certain thresholds, for example reporting all ages above 90 as “90 or more;” swapping data values for selected records, e.g., exchange two or more individuals’ demographic data (Dalenius and Reiss, 1982); and, adding noise to numerical data
values (Brand, 2002; Yancey et al., 2002). See Willenborg and de Waal (2001), Working Paper 22 of the Federal Committee on Statistical Methodology (2005), and Duncan and Stokes (2009) for general overviews of statistical disclosure limitation techniques.

As the amount of information that is readily available to the public continues to expand, e.g., via the Internet and private companies, these techniques may have to be applied with high intensity to ensure adequate protection. However, applying these methods with high intensity can have serious consequences for secondary statistical analyses. For example, aggregation of geography to high levels disables small area estimation and hides spatial variation; top-coding eliminates learning about tails of distributions—which are often most interesting—and degrades analyses reliant on entire distributions (Kennickell and Lane, 2006); swapping at high rates destroys correlations among swapped and not swapped variables (Winkler, 2007; Drechsler and Reiter, 2010b); and, adding random noise introduces measurement error that distorts distributions and attenuates correlations (Fuller, 1993). In fact, Elliott and Purdam (2007) use the public use files from the UK census to show empirically that the quality of statistical analyses can be degraded even when using recoding, swapping, or stochastic perturbation at modest intensity levels. These problems would only get worse with high intensity applications.

This article reviews an approach to releasing highly redacted, public use data that theoretically can avoid many of these problems. The idea, generally referred to as synthetic data, is to replace confidential data with draws from statistical models, and release these (partially) simulated datasets. The article focuses on the two main variants of synthetic data: full synthesis (every released value is simulated) and partial synthesis (some values are simulated, and others are original). The article summarizes typical approaches to data generation and disclosure risk assessment for synthetic data. The concluding section lays out some research challenges in the topic. In what follows, the organization releasing synthetic public use data is abbreviated as the imputer, and the ill-intentioned user seeking disclosures is abbreviated as the intruder.

Overview of full and partial synthesis

Fully synthetic data

Fully synthetic data was first proposed by Rubin (1993), who suggested that the imputer (i) randomly and independently sample units from the sampling frame to comprise each synthetic data set, (ii) impute the unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) release multiple versions of these datasets to the public. A similar proposal was put forth by Fienberg (1994). Methods for obtaining valid inferences from multiple synthetic datasets were derived by Raghunathan et al. (2003) and Reiter (2005c). Reiter (2002) illustrates the impact of the sampling design and the number and size of synthetic datasets on inferences.

To illustrate how fully synthetic data might work in practice, we modify the setting described by Reiter (2004a). Suppose the imputer has collected data on a random
sample of 10,000 people. The data comprise each person’s race, sex, income, and indicator for the presence of a disease. We assume the imputer 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 distribution. We assume the imputer knows the income and disease status only for the people who respond to the survey.

To generate synthetic data, first the imputer randomly samples some number of people, say 20,000, from the population list. The imputer then generates values of income and disease status for these 20,000 synthetic people by randomly simulating values from the joint (posterior predictive) distributions of income and disease status, conditional on their race and sex values. These distributions are estimated using the collected data and possibly other relevant information. The result is one synthetic data set. The imputer repeats this process say ten times, each time using different random samples of 20,000 people, to generate ten synthetic datasets. These ten datasets are then released to the public.

When the incomes and disease statuses for the synthetic people are simulated from the true joint probability distributions, the synthetic data should have similar characteristics on average as the collected data. The on average caveat is important: parameter estimates from any one simulated data set are unlikely to equal exactly those from the observed data. The synthetic parameter estimates are subject to three sources of variation, namely (i) sampling the collected data; (ii) sampling the synthetic units from the population; and, (iii) 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 the multiple imputation framework applies. The analyst estimates parameters and their variances in each of the synthetic datasets, and combines these results using the methods of Raghunathan et al. (2003).

Fully synthetic data can have positive data utility features. When data are simulated from distributions that reflect the distributions of the observed data, frequency-valid inferences can be obtained from the multiple synthetic datasets for a wide range of estimands. These inferences can be determined by combining standard likelihood-based or survey-weighted estimates; the analyst need not learn new statistical methods or software programs to account for the disclosure limitation. 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. The released data can include simulated values in the tails of distributions (no top-coding) and avoid category collapsing. 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 (Reiter, 2005b).
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 imputers need to release information that helps analysts decide whether or not the synthetic data are reliable for their analyses. See the conclusion for further discussion of this issue.

**Partially synthetic data**

First proposed by Little (1993), partially synthetic data comprise the units originally surveyed with some collected values replaced with multiple imputations. The imputations are drawn from distributions designed to preserve important relationships in the confidential data. To illustrate a partial synthesis, suppose the imputer wants to replace income when it exceeds $100,000 and is willing to release all other values. The imputer generates replacement values for the incomes over $100,000 by randomly simulating from the distribution of income conditional on race, sex, and disease status. To avoid bias, this distribution also must be conditional on income exceeding $100,000. The distribution is estimated using the collected data and possibly other relevant information. The result is one synthetic dataset. The imputer repeats this process say ten times to generate ten synthetic datasets, which are released to the public.

Partial synthesis can maintain many of the benefits of full synthesis, in that the released data look like the original data (no top-coding, collapsing, etc.). Inferences remain straightforward: the analyst estimates parameters and their variances in each of the synthetic datasets, and combines these results using the methods of Reiter (2003). An advantage of partial synthesis relative to full synthesis is that only a fraction of the data are imputed, so that analysts’ inferences are generally less sensitive to the imputer’s model specification. Unlike fully synthetic data, partially synthetic data must be analyzed in accordance with the original sampling design. Missing values can be imputed at the same time as replacement values using the approach in Reiter (2004b).

With partial synthesis, the imputer needs to determine which values to synthesize. Most current applications (listed below) involve synthesis of entire variables deemed by the imputer to be potentially available to intruders. However, synthesis of entire variables is not necessary: synthesis can be targeted to particular values for at-risk records. For example, Drechsler and Reiter (2010b) begin with synthesis of entire variables, and add back genuine values in a manner that does not significantly increase the disclosure risks.

Partially synthetic data products have been developed by several national statistical agencies in the U.S. The Federal Reserve Board in the Survey of Consumer Finances replaces monetary values at high disclosure risk with multiple imputations, releasing a mixture of these imputed values and the unreplaced, collected values (Kennickell, 1997). The Census Bureau has released a partially synthetic, public use file for the Survey of Income and Program Participation (SIPP) that includes imputed values of
Social Security benefits information and dozens of other highly sensitive variables (Abowd et al., 2006). The Census Bureau protects the identities of people in group quarters (e.g., prisons, shelters) in the American Community Survey by replacing demographic data for records at high disclosure risk with imputations (Hawala, 2008). The Census Bureau also has developed synthesized origin-destination matrices, i.e. where people live and work, available to the public as maps via the web (Machanavajjhala et al., 2008). Partially synthetic, public use datasets are being developed for the Longitudinal Business Database (Kinney and Reiter, 2007) and the Longitudinal Employer-Household Dynamics database. Other applications of partially synthetic include Abowd and Woodcock (2001, 2004), Little et al. (2004), Graham and Penny (2005), An and Little (2007), Drechsler et al. (2008a,b), and Graham et al. (2009).

Data Generation and Risk Assessment

Data generation

The key to the success of synthetic data approaches, especially when replacing many values, is the data generation model. Current practice for generating synthetic data typically employs sequential modeling strategies based on parametric models similar to those for imputation of missing data in Raghunathan et al. (2001). The basic idea is to impute $Y_1$ from a regression of $Y_1$ on $(Y_2, Y_3, etc.)$, impute $Y_2$ from a regression of $Y_2$ on $(Y_1, Y_3, etc.)$, impute $Y_3$ from a regression of $Y_3$ on $(Y_1, Y_2, etc.)$, and so on. An advantage of this strategy is that it is generally easier to specify plausible conditional models than plausible joint distributions. A disadvantage is that the collection of conditional distributions is not guaranteed to correspond to a proper joint distribution.

Specifying these conditional models can be a daunting task: typical data include numerical, categorical, and mixed variables, some of which are not easy to model with standard tools, and relationships among these variables can be nonlinear and involve interactions. To reduce reliance on parametric assumptions, several authors have investigated semiparametric and nonparametric approaches to data generation, including generalized additive models (Raghunathan, 2003), classification and regression trees (Reiter, 2005d), kernel density regressions (Woodcock and Benedetto, 2009), Bayesian networks (Young et al., 2009), random forests (Caiola and Reiter, 2010), and support vector machines (Drechsler, 2010). A survey of the literature suggests that these methods can generate reasonably high-quality data with minimal agency tuning; however, empirical results clearly indicate that there is much room for improvement on the data quality dimension (Drechsler and Reiter, 2010a).

Risk Assessment

Releasing fully synthetic data makes identification of units and their sensitive data difficult. 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. Fully synthetic data are not risk free, however. For example, it may be highly unlikely to simulate a certain rare combination of variables unless a particular record is in the original data. An intruder may infer the presence of the particular record when the synthetic datasets contain that rare combination.

The protection afforded by partial synthesis depends on the nature of the synthesis. Replacing key identifiers with imputations makes it difficult for intruders to know the original values of those identifiers, which reduces the chance of identifications. Replacing values of sensitive variables makes it difficult for users to learn the exact values of those variables, which can prevent attribute disclosures. Nonetheless, there remain disclosure risks in partially synthetic data no matter which values are replaced. Analysts can utilize the released, unaltered values to facilitate disclosure attacks, for example via matching to external databases, or they may be able to estimate genuine values from the synthetic data with reasonable accuracy.

Thus far, the synthetic data products released by statistical agencies have relied on partial synthesis. For these products, imputers have used three approaches to quantify disclosure risks. The first is to use record linkage techniques to attempt to match records in the original data and each synthetic dataset. The risk metric is the overall percentage of correct matches. This was done, for example, to assess disclosure protection in the SIPP synthetic data (Abowd et al., 2006). Re-identification experiments have the benefit of being closely tied to what intruders might actually do; however, they have several weaknesses. First, they make the rather strong assumption that the intruder has access to the entire original data file, which may be unrealistic in some cases. Second, they do not account for the protection engendered by the process of sampling, which is known to reduce disclosure risks. Third, they do not utilize any released information about the synthesis models to assist the record linkage.

A second approach is to compute probabilities of identification, e.g., using methods developed by Reiter and Mitra (2009) and Drechsler and Reiter (2008). Related approaches are described by Duncan and Lambert (1989), Fienberg et al. (1997), and Reiter (2005a). To describe these, suppose the intruder has a vector of information, t, on a particular target unit in some population D of size N. Let t0 be the unique identifier of the target, and let Dj0 be the (not released) unique identifier for record j in the synthetic datasets Dsyn. Let S be any information released about the simulation models. The intruder’s goal is to match unit j in Dsyn to the target when Dj0 = t0. Let J be a random variable that equals j when Dj0 = t0 for j ∈ Dsyn. The intruder thus seeks to calculate the Pr(J = j|t, Dsyn, S) for j = 1, . . . , N. Disclosure risk can be quantified with summaries of these identification probabilities; for example, the number of times the record j with the highest value of Pr(J = j|t, Dsyn, S), if a unique maximum exists, is in fact t. Some benefits of probabilities of identification include (i) they can account for specific intruder information, including whether or not the intruder knows a record is in the sample; (ii) they can account for information released about the synthesis design; and, (iii) they can account formally for the number of released datasets. However, they have the drawback that they require the imputer to posit possible sets of information known by the intruder. In general, it is impossible
for imputers to know this information, so that the best they can do is to assess disclosure risks under different threat scenarios. Probabilities of identification also are more computationally complex than record linkage experiments.

The third approach, used in the On The Map application (Machanavajjhala et al., 2008), is to measure the amount of protection according to the risk criterion differential privacy (Dwork, 2006; Abowd and Vilhuber, 2008). This criterion bounds the amount of information that an intruder can learn about any aspect of the original database regardless of the nature of the prior information held by the intruder. For example, it protects against an intruder who knows the identities and attribute values of every record in the database except for one record. This is a great strength of the differential privacy framework, in that the imputer need not guess at the intruder’s prior information. However, differential privacy criteria have some disadvantages, including (i) they do not account for sampling; (ii) they can result in metrics that are difficult to interpret; (iii) they may require high amounts of data distortion since the imputer must guard against all possible sets of prior information; and, (iv) they are hard to compute with complex, high dimensional data. Indeed, thus far, in practice differential privacy risk has been computed only for synthesizers based on multinomial likelihoods with Dirichlet prior distributions.

**Conclusion**

For both fully and partially synthetic data, the main challenge is specifying imputation models that give valid results. More research is needed on flexible approaches to data generation, for example based on nonparametric Bayesian models.

Even with nonparametric synthesis methods, undoubtedly some synthetic data inferences will deteriorate significantly because of imperfect imputation models. It is arguably essential that agencies develop ways to provide feedback to users about the quality of inferences from synthetic data—or any other disclosure protection method applied with high intensity—for specific estimands. One possibility is to build a verification server, as suggested by Reiter et al. (2009). The basic idea is as follows. The synthetic data user submits a description of the analysis to the verification server; for example, regress attribute 5 on attributes 1, 2, and 4. The verification server performs the analysis on both the confidential and synthetic data, and from the results calculates analysis-specific measures of the fidelity of the one to the other. For example, for any regression coefficient, measure the overlap in its confidence intervals (Karr et al., 2006) computed from the confidential and synthetic data. The verification server returns the value of the fidelity measure to the user. If the user feels that the intervals overlap adequately, the synthetic data have high utility for that analysis. With such feedback, analysts can avoid publishing—in the broad sense—results with poor quality, and be confident about results with good quality. Unfortunately, fidelity measures also leak information about the confidential data that could be used for disclosure attacks (Reiter et al., 2009). Hence, despite the clear benefit of verification servers, it is an idea without practicable methodology to date.
References


**Cross-References**

Data confidentiality, Sample survey, Data masking for disclosure limitation