RESEARCH REPORT SERIES
(Disclosure Avoidance #2017-01)

Philosophy of Disclosure Avoidance for Census Bureau Data

Michael H. Freiman

Center for Disclosure Avoidance Research
U.S. Census Bureau
Washington DC 20233

Report Issued: June 13, 2017

Abstract

The Census Bureau is mandated by law to collect data and publish statistical summaries. In so doing, the Bureau must ensure that the data are used solely for statistical purposes and that the privacy and confidentiality of responding individuals and organizations are not compromised by any publications. The law does not mandate how these two requirements are balanced when tradeoffs are necessary. Methods used by the Census Bureau to protect privacy and confidentiality are designed to guard against unauthorized identity, attribute, or inferential disclosure. Current disclosure methods used by the Bureau include suppression, aggregation/coarsening, perturbation by input or output noise, and the creation and release of synthetic data. However, each of these methods invariably involves a tradeoff between privacy loss and accuracy: the more accurate the data, the more privacy that is lost. Legacy disclosure avoidance systems did not quantify either the privacy loss or the accuracy of the resulting data. The Census Bureau is now moving to a new generation of disclosure avoidance techniques based on formal privacy methods that quantify both of these measures and allow policymakers to specify the tradeoff between privacy and data accuracy. The changes in disclosure limitation methodology applied to the Census Bureau data may result in a larger group of researchers who realize that the public versions of released data and statistics are not suitable for their research. As a result, it is possible that more researchers will request access to the Federal Statistical Research Data Centers in the future.

Key Words: Confidentiality, Database Reconstruction, Formal Privacy, Reidentification, Synthetic Data

(This page intentionally left blank)

Philosophy of Disclosure Avoidance for Census Bureau Data

Michael Freiman

U.S. Census Bureau, Washington DC, United States

The Census Bureau is mandated by law to collect and release data, while ensuring that the data are used solely for statistical purposes and that the privacy and confidentiality of responding individuals and organizations are not compromised in any way. Specifically, the Census Bureau and its agents may not:

(1) "Use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or
(2) Make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or
(3) Permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports." (Title 13, U.S. Code, Section 9.)

Similar restrictions are in place for Federal Tax Information (Title 26, U.S. Code, Section 6103), which the Census Bureau also uses to accomplish its mission.

Although the law mandates that the Census Bureau must both publish and protect data, it does not indicate how to balance these two requirements when tradeoffs are necessary.

The Census Bureau must protect against three main types of unauthorized disclosure:

- **Identity disclosure** (reidentification): manipulation of released data reveals the identity of an individual or business.
- **Attribute disclosure**: manipulation of the data reveals some feature of a respondent. An attribute disclosure may occur in addition to an identity disclosure or on its own.
- **Inferential disclosure**: the data user can determine an identity or attribute with a high probability. An inferential disclosure occurs when the user's posterior belief regarding the particular record differs substantially from the user's prior belief.

Identity and attribute disclosure are special cases of inferential disclosure where the posterior belief is unity. Controlling inferential disclosure is, therefore, the general form of disclosure avoidance.

When traditional disclosure avoidance methods were developed, the risk of disclosure was much lower for any given data release than it is today because in the past, there were relatively few external data available to correlate with any given data or statistical release and the database reconstruction theorem (Dinur and Nissim, 2003) was unknown. The Census Bureau and other agencies curated most of the data that could compromise confidentiality, and the few commercial and private exceptions were mostly known and appropriately addressed. Database reconstruction attacks were completely unknown, and historical products were not designed to counter such attacks.

More recently, the amount of publicly available or proprietary data that can create a disclosure risk has greatly increased. "Big data" has become a crucial asset for businesses, which use the data to tailor their products. Simultaneously, research has improved the algorithms for attack, and increased computing power allows the data and algorithms to be put into practice, increasing the risk of a disclosure dramatically.

The advent of "big data" has made the consequences of a database reconstruction attack more serious. A database reconstruction uses only the information actually published by the statistical agency to create a record-level image of the confidential data that includes every tabulation variable used in every published table, regression analysis, or any other released output. This means that the cumulative publication system must be protected, not just each individual release. If there is already a four-digit NAICS table at the county level for a particular analysis variable, even though a new analysis doesn't publish the NAICS coefficients, the published regression coefficients controlling for those effects inform the reconstruction. This is not just a curiosity. The Census Bureau published at least 5.4 billion independent statistics on 308 million people in the 2010 Census of Population and Housing in the redistricting files and Summary File 1 alone.

The more accurate the database reconstruction possible, the more likely a successful reidentification becomes.

Types of Disclosure

The most obvious type of disclosure is reidentification, and several high-profile reidentifications have already occurred. In 2006, Netflix released an anonymized dataset listing movie ratings from over 400,000 users and the dates they were given, using a special identifying code for each user that Netflix did not intend to be traceable to a person in the real world. Two years later, Narayanan and Shmatikov (2008) described an algorithm to determine with high certainty that a given user in the Netflix dataset is the same as a given rater on the Internet Movie Database (IMDb), where movie ratings and often raters' identities are public. Thus Narayanan and Shmatikov identified some of the Netflix users. They proposed that for 99% of Netflix users, the user's record could be uniquely identified in the dataset if an intruder had eight movie ratings and the dates they were made, even assuming that only six of the intruder's ratings are correct and that the dates might be off by up to two weeks. The authors pointed out that IMDb raters are not the only ones at risk; anyone who mentions movie likes or dislikes in a blog or in conversation could potentially be linked to the Netflix data. The Narayanan and Shmatikov attack is an example of successful reidentification that was enabled by a very accurate database reconstruction. Netflix had not used formally private methods to anonymize the data it released. Narayanan and Shmatikov exploited this to reconstruct which IMDb users where in the Netflix data.

In another instance, the Massachusetts Group Insurance Commission released data to researchers with obvious identifying information (e.g., name and address) removed. Sweeney (2002) identified the medical records of then-Governor William Weld, matching them with cheaply purchased voter registration rolls from Cambridge, Massachusetts, where Weld resided, which included the voter's sex, date of birth and 5-digit ZIP code. Eleven years later, Sweeney (2013) showed that information in newspaper articles about accidents could be used to reidentify 35 of 81 records in an anonymized health records dataset released by the state of Washington, noting that "employers,

financial organizations and others" would have access to the information in the newspaper accounts.

Types of Disclosure Protection

Current disclosure methods fall into several categories:

- **Suppression.** The data release omits some values.
- **Aggregation/coarsening.** The release gives values in limited geographic or topical detail.
- **Perturbation by input or output noise.** Modified data are published. The releasing agency may add noise to input data, subject them to data swapping (a form of input noise), round them, add noise to the output, or apply some other method.
- **Synthetic data.** Some or all of the originally collected data are replaced by data generated using a model, intended to have similar properties to the original data.

We believe that while most of these current methods afford substantial protection, they have shortcomings that are becoming more significant with the passage of time. The problem with the current methods is that they are ad hoc. There is no principled definition of the global disclosure risk. Therefore, it is not possible to quantify the extent to which a particular data release compromises the confidentiality of the underlying data in light of what has already been published.

Data providers may be confident they have protected against certain types of attacks, but without this global risk measure, they cannot ever quantify that protection. Similarly, ad hoc disclosure limitation methods cannot determine which future attacks have been defended. Improved database reconstruction algorithms make these disclosure risks even more pronounced.

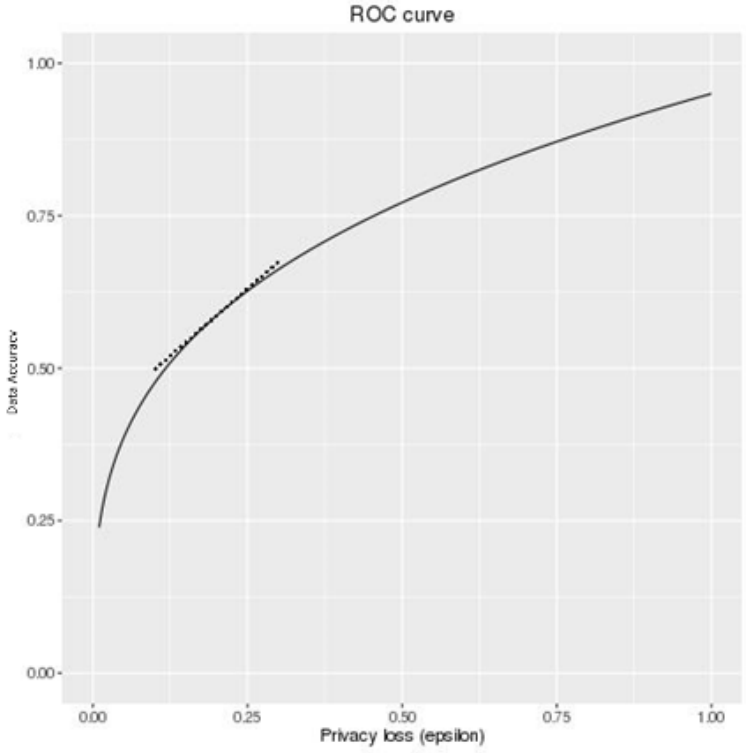
Another problem with current methods is a lack of transparency. In most cases, the data releasing agency does not reveal the disclosure parameters that were used to make the data safe for release. Such parameters typically include the swapping rate or the amount of noise that was added. As a result, data users cannot address the amount of noise, bias or error that might have been introduced into their work products by the disclosure protection. This represents a problem for both external validity and reproducibility.

Theoretical Considerations

Disclosure protection faces a bigger challenge: all data publication leads to some privacy loss. This is the fundamental consequence of the database reconstruction theorem. No one line separates protecting privacy and not protecting privacy. Privacy loss is often incremental across several data releases, any one of which may have minimal risk (Abowd, 2016). Every piece of output leads to some actual privacy loss, albeit sometimes small, if only in the form of causing slightly more accurate inference about individual records in the dataset. We must consider the risk incurred by releasing even a little more data in the context of data already released. If privacy loss and data a are quantified, we may evaluate the tradeoff between the two in a Receiver Operator Characteristics (ROC) curve. The curve below shows an example of one possible frontier of possible amounts of privacy loss and data accuracy. An agency should protect data up to the point where the marginal privacy loss equals the marginal willingness to incur privacy loss both stated in terms of the incremental data accuracy from the privacy loss. These are shown in this illustration. The ROC

curve (or Production Possibilities Frontier to economists) is the technology of disclosure avoidance. Every point on the ROC curve represents a feasible disclosure avoidance system. Those closer to the origin involve less privacy loss and less data accuracy. Those farther from the origin involve more privacy loss and data accuracy. Publicizing the confidential microdata would be represented by infinite privacy loss and the most accurate possible data.

The social choice a statistical agency faces is where to locate on the ROC. Every point is a legitimate disclosure avoidance system. One way to represent the preferences over data accuracy and privacy loss is by parallel lines like the one shown by the dotted tangent line in the figure. The slope of the tangent line, and hence the point that best balances data accuracy and data privacy, is a policy decision. The Census Bureau is trying to measure that balance, but such measurements do not depend on properties of the disclosure avoidance system or the data generating mechanism. They represent measurements of social welfare very similar to the social welfare analysis that underlies environmental or regulatory policy.



Several theorems demonstrate the challenges of keeping data private. The Database Reconstruction Theorem (Dinur and Nissim, 2003) shows that any finite database may be reconstructed arbitrarily accurately using finitely many queries. Hence, we must limit the amount and precision of output. Dwork *et al.* (2006) and Dwork (2006) show that controlling the accuracy of a reconstructed database to within provable limits requires noise infusion with particular properties. They prove these properties hold for the formally private disclosure avoidance system known as differential privacy. Dwork and Naor (2010) prove that the only way to make an inferential disclosure impossible is to publish no data or fully encrypted data.

The Census Bureau is moving from ad hoc disclosure avoidance methods toward new methods based on a formal definition of privacy. Using these methods, the amount of privacy loss can be

quantified and limited. The method used guarantees this level of privacy, and this guarantee can be mathematically proven. A formal privacy criterion refers to the guarantee, which holds regardless of an intruder's prior knowledge, rather than to the method used to fulfill the guarantee. In 2008, the Census Bureau released OnTheMap, a web-based mapping and reporting application that shows where workers are employed and where they live. This was the first real-world production deployment of such formally private methods (Machanavajjhala, *et al*., 2008).

The best known formal privacy criterion is differential privacy (Dwork *et al*., 2006). This criterion envisions a user who receives output based on a dataset via a privacy-preserving publication algorithm. To meet the criterion, the Bayes factor that transforms the intruder's prior probabilistic belief about any individual record into posterior beliefs must be approximately the same regardless of whether the algorithm uses the full dataset or the dataset with one record omitted for any potential input dataset. ("Approximately the same" is formally defined, but we do not go into the technical details here.) The guarantee must hold true even if the intruder knows the value of every variable for every record in the dataset other than the aforementioned single record. In releasing data under this or another formal privacy criterion, the releasing agency creates a privacy-loss budget, indicating how much cumulative loss of privacy is allowable for the dataset across all queries. Formal privacy also allows more openness than most current methods, as all parameters of the disclosure protection algorithm are released and known, unlike the current paradigm in which thresholds and the swapping rate are confidential.

The Future of Disclosure Avoidance at the Census Bureau

For some data products, the Census Bureau is now researching the use of synthetic data that could be released in place of the original data. Synthetic data would be generated from a model, itself based on the original data. Hence, the synthetic data retain many properties of the original data. Synthetic data are only as good as the model used to generate them and can never reflect all properties of the original data. Hence, "only some hypotheses can be studied accurately" (Abowd and Schmutte, 2015). Since results may be somewhat distorted, a synthetic dataset should ideally be accompanied by the opportunity to check one's work using a *verification server,* such as the ones described in Reiter *et al.* (2009). The verification server performs the researcher's analysis on these original data and provides a metric of how similar the results are between the two datasets. If the results are similar, a researcher might view the research as complete and valid, while if the results are substantially different, the researcher might view the results with skepticism or obtain access to the restricted data—for example, at a Federal Statistical Research Data Center (FSRDC)—to rerun the analysis. This approach can lead to fewer privacy concerns than traditional methods, but there is still some privacy loss from running the output through the validation server.

The Census Bureau will continue to research new methods of making data both accurate and private. Much current work focuses on provably-safe microdata, a form of synthetic data with formal privacy guarantees and provable accuracy for pre-specified analyses. Ultimately, we hope to have data that are formally private, although we may need several iterations of disclosure avoidance modernization to accomplish this.

Outlook

The changes in disclosure limitation methodology applied to Census Bureau data may result in a larger group of researchers who realize that the public versions of the data are not suitable for their

research. This may stimulate the need for more verification servers or other approaches for accessing confidential data in a mathematically safe manner. It is quite possible that more researchers will request access to the FSRDCs to complete their research in the future, especially when the validity of their original results is called into question by the verification server.

It will take some time for most Census Bureau data products to incorporate these new methods. For now, Census Bureau datasets will continue to rely heavily on traditional methods. However, due to the continually increasing disclosure risk, we have recently changed some of our disclosure avoidance practices to mitigate the risk appropriately and will continue to reevaluate them. In so doing, we consider the tradeoff between data accuracy and disclosure risk. For instance, we now enforce rounding rules for most counts and estimates. In most cases, releasing estimates, say, to seven significant digits instead of four does not add enough extra accuracy to warrant the additional disclosure risk. We are also extra mindful that the overall quantity of output is commensurate with what researchers need to fulfill their approved research projects.

References

Abowd, J. M. (2016). How Will Statistical Agencies Operate When All Data Are Private? Available at http://digitalcommons.ilr.cornell.edu/cgi/viewcontent.cgi?article=1029&context=ldi, forthcoming, *Journal of Privacy and Confidentiality*.

Abowd, J. M., & Schmutte, I. M. (2015). Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity*, *2015*(1), 221-293.

Dinur, I., & Nissim, K. (2003, June). Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 202-210). ACM.

Dwork, C. (2006). Differential privacy. *Automata, languages and programming*, 1-12.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006, March). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference* (pp. 265-284). Springer Berlin Heidelberg.

Dwork, C., & Naor, M. (2010). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy, *Journal of Privacy and Confidentiality*. Available at http://repository.cmu.edu/jpc/vol2/iss1/8/.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008, April). Privacy: Theory meets practice on the map. In *Data Engineering, 2008. ICDE 2008* (pp. 277-286). IEEE.

Narayanan, A., & Shmatikov, V. (2006). How to break anonymity of the Netflix prize dataset. *arXiv preprint cs/0610105*.

Reiter, J. P., Oganian, A., & Karr, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics & Data Analysis*, *53*(4), 1475-1482.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557-570.

Sweeney, L. (2013). Matching known patients to health records in Washington State data. Data Privacy Lab. 1089-1. Available at https://dataprivacylab.org/projects/wa/1089-1.pdf.

Title 13, U.S. Code, Section 9. Available at https://www.gpo.gov/fdsys/pkg/USCODE-2009-title13/html/USCODE-2009-title13.htm.

Title 26, U.S. Code, Section 6103. Available at https://www.gpo.gov/fdsys/pkg/USCODE-2009-title26/html/USCODE-2009-title26.htm.