An Economist's Guide to Statistical Disclosure Limitation Decision-making

John M. Abowd

January 2023

Abstract

Incomplete draft. Essentially a long outline.

1 Background

In our 2015 Brookings paper (Abowd and Schmutte, 2015a), Ian Schmutte and I illustrated the pitfalls for inference validity when social scientists fail to account for the statistical disclosure limitation (SDL) treatments applied to data released to the public. In that work, we noted that SDL has many similarities to missing data when it is applied directly to the confidential data before the analyst has determined whether the hypothesis under study is estimable using the original data. In particular, SDL can cause nonignorable missing data, particularly when the method is suppression based on the unpublished statistic's value. We provided the full Bayesian solution to the inference problem when the SDL is known or discoverable, regrettably in the online appendix that BPEA no longer curates (but see Abowd and Schmutte (2015b)). We illustrated the problems using several designs common in economics, some of which I will also use in this paper.

In our 2019 AER paper (Abowd and Schmutte, 2019), we developed a complete economic framework for analyzing confidentiality protection implementations and choosing their parameters. Ours was a very stylized model, but it made the important point that SDL frameworks are technologies that, if properly implemented, describe a production possibilities set with a frontier (PPF) that quantifies the trade-off between the accuracy and confidentiality parameters. More importantly, we noted that the social choice problem requires a method of aggregating heterogeneous preferences over those parameters, and that the computer science underlying the PPF does not provide any tools for specifying the social welfare function. That paper collapsed the decision problem into two dimensions: a single disclosure risk parameter and a single measure of accuracy, called "utility" in most of the SDL literature, but more properly labeled "fitness-for-use" because the utility of the SDL system depends jointly on confidentiality risk and accuracy. While multidimensional representations are more difficult to illustrate, current discussions make clear that a more complete analysis requires

specifying multiple fitness-for-use and privacy-protection dimensions. In particular, two target use cases imply at least two dimensions for accuracy and another two for controlling the privacy-loss (total and allocation to each use case).

As I noted (with others) in Abowd *et al.* (2019), it is important for economists and other social scientists to participate in the sea change that is occurring in SDL frameworks. Early contributions include (Duncan *et al.*, 2001) and (Heffetz and Ligett, 2014). Recently, other social scientists and statisticians have taken up the challenge Slavković and Seeman (2023); Gong and Meng (2020); Hotz *et al.* (2022). This conference is an opportunity to move that discussion forward.

In this paper, I lay out some of the issues that economists have, so far, struggled to address. I draw heavily on Jarmin *et al.* (n.d.), a response to Hotz *et al.* (2022), for characterizing properties of SDL frameworks. I extend the analysis in Abowd and Schmutte (2015a, 2019) and illustrate the extension in the context of several competing use cases. My goal is to encourage economists, and social scientists generally to set the problem up properly and, hence, offer useful guidance to data curators, particularly statistical agencies. In Section 2, I illustrate the difference between the researcher's problem and the statistical agency's. In Section 3, I address the problem of analyzing arbitrarily many studies from a single confidential source. Section 4 considers the equally-vexing problem of pre-specifying all hypotheses. In Section 5, I show how to integrate the analysis of multiple research objectives with the principles of SDL frameworks, then, in Section 6 illustrate this integration by considering two competing research uses within the social welfare maximizing framework. Finally, in Section 7, I make some recommendations for future research.

2 One researcher's problem is not the same as the statistical agency's problem

Researchers usually consider the application of SDL in one of two scenarios: (a) the only release from a particular confidential source is their particular study; and (b) the SDL has been previously applied and the only inputs available for their particular study are the published data from such application. Either of these scenarios can be useful, but neither is a complete statement of the statistical agency's problem.

Considering the researcher's study as the only one to be published from the confidential data is inconsistent with most statistical agency mandates to collect data for publication and use by government agencies, businesses, and the general public.

Considering the SDL as given is scientifically inappropriate if the design of the original data supports the the inference under study. In that case, the salient questions are: (1) does there exist a confidentiality protection framework under which the agency can agree to publish the results of the study? and (2) is that confidentiality protection framework compatible with the other publications planned or already conducted on the confidential data?

Either case requires SDL frameworks that generate global disclosure risk assessments covering both the researcher's proposed study and all other planned or ad hoc publications. If publications are to be released sequentially, then that framework must also measure the

incremental disclosure risk associated with the next publication, given those already released or scheduled for release, or control global risk in some other well-defined manner.

The only traditional SDL method that provides both a global disclosure limitation measure and a proof that sequential publications continue to meet it is the one used by the U.S. Census Bureau for tabular summaries of economic data. This method is known as primary/complementary suppression. Cox (1995) proved that the solution to a particular network integer program, if it exists, produces complementary suppressions that guarantee that all tables produced from the confidential input safeguard the primary suppressions of every table. This permits sequential release of new tables as long as they are processed by the correct network integer program extension of the solution that yielded the currently published tables. This principled suppression method measures the global disclosure risk with a rule, called the primary suppression rule, that sets a limit to the precision that an attacker can achieve in estimating the specific value of the feature of any entity found in in the confidential database. For example, the inference on the reported value (in the survey) of a particular business's sales should be no more accurate than $\pm 30\%$. Then, it accumulates complementary suppressions that safeguard the primary suppressions in all tables as they are cumulatively processed for release.

Primary/complementary suppression it is completely rigid. Once the data curator establishs the global risk measure, there is no role for uncertainty in the implementation. A cell or statistic is either releasable or not, and the order in which the agency askes the question matters because this method does not satisfy the second property above—it does not produce a measure of incremental disclosure risk. The reason is that primary/complementary suppression does not compose (see Section 5.1). If the system is applied independently to two different data products from the same confidential source (or confidential sources with significant overlap; e.g., Census and BLS business data), there is no proof that the primary suppressions in either data publication remain protected. And, in general, they do not.

Other traditional SDL frameworks simply ignore the problem of ensuring that two publications from the same confidential input can be safely released together. The most egregious example of this failure is the use of aggregation to produce tabular summaries at one level of geography or industry and coarsening over the same geographic or industrial variables to produce regression estimates or microdata inputs for other analyses. Those SDL methods do not compose (i.e., they do not provide a measure of the incremental risk of each product such that the sum of these incremental risks is upper-bounded by the global risk measure). They expose the confidential data because the statistics in one release can be use to undo the protections in the other—an example of post-processing vulnerability (see Section 5.2.

Differential privacy frameworks solve these problems, but, as importantly, the methods used to establish the properties of differentially private frameworks expose the vulnerabilities of alternative ways of conceptualizing confidentiality protection. The most important of these vulnerabilities is the impossibility theorems found in Dinur and Nissim (2003); Dwork and Naor (2010). It is impossible to define an absolute-risk confidentiality protection framework (i.e., one that measures the probability that a confidential datum will be exposed by a particular publication) that simultaneously defends against an arbitrary attacker and allows valid generalizable inference from the confidential data. There is a simple example in Abowd and Hawes (2023). The point of the disclosure risk measure that Dwork (2006) proposed as a solution to the impossibility of controlling absolute disclosure risk is to compare the

inference with and without any particular entity's data and for all possible entities and configurations of data. Applying these principles yields SDL frameworks that can still measure any disclosure risk summary that matters to users and also protects scientific inferences from undue influence of one or a few observations in a manner comparable to robust statistics as Dwork and Lei (2009); Wasserman and Zhou (2010); Kifer *et al.* (2022) elaborate.

The statistical agency is, therefore, confronted with a much different problem than the one social scientists usually address. The data steward must insure that all publications from the ensemble of confidential inputs in its custody cumulatively protect the confidentiality of the underlying data. The data steward must therefore choose an SDL framework that accommodates all anticipated publications or analysis types and accept the loss of precision, or even identification, in some potential publications as a cost of releasing others.

3 The fallacy of supporting arbitrarily many analyses from the same confidential data

Dinur and Nissim Dinur and Nissim (2003) proved that publishing too many statistics too accurately exposes the records of the confidential database, in the schema used for publication, within a finite number of published statistics with probability arbitrarily close to unity. This result, which predates formally private SDL frameworks like pure differential privacy, depends only on the algebra of forming statistical tables from record-level data (Garfinkel *et al.*, 2018). The only way to support arbitrarily many analyses from the same confidential data is to release those data unaltered. Whether such an action is ever consistent with an agency's confidentiality protection framework, the only analyses that will be feasible are those supported by the schema of the released data.

4 The fallacy that all hypotheses must be pre-specified

A researcher's basic requirement for an SDL framework is that if the original data generating mechanism can support the inference under study, then the SDL framework should also. It is incumbent upon the data curator to control the disclosure risk through mechanisms that implement its chosen SDL framework and are consistent with the requirements to do global risk analysis laid out in Section 5. It is incumbent on the researcher to demonstrate that a particular research design, incorporating mechanisms that can implement the curator's SDL framework, delivers inferences of known quality. Sometimes, this will not be possible. Such impossibility results are important to catalogue, since they should guide the choice of SDL frameworks if they apply to a class of problems that the data curator wishes to support.

5 Integrating the researcher's problem into the custodian's SDL framework

This section will illustrate the importance of composition, resistance to post-processing decay, and inference validity in the design of SDL frameworks.

5.1 The composition requirement

Mechanisms implementing an SDL framework compose if the global risk measure associated with the ensemble of data publications is upper-bounded by the sum of the risk analyses for each product considered individually. Many SDL frameworks fail composition. Sometimes the implementing mechanism can be modified to ensure that global disclosure risk is controlled. Otherwise, frameworks that do not compose prevent the use of social welfare maximizing decision-making.

5.2 The post-processing requirement

SDL frameworks that resist post-processing decay use mechanisms that cannot be undone after the data are released. The risk that a confidentiality mechanism can be circumvented by post-processing increases as the number of statistics released from the source increases and is compounded if the implementing mechanisms do not compose. Since a statistical agency must adjust its dissemination strategy if there is post-processing risk, it is important to incorporate these adjustments–mostly implemented in what is known as tier-access systems, into the framework.

5.3 The inference validity requirement

Inference validity means that the SDL framework supports inferences in the sense defined as "known or discoverable SDL" from (Abowd and Schmutte, 2015a). Some SDL implementations may not be "congenial" Gong and Meng (2020) to certain hypotheses; however, there often exist mechanisms that support the same SDL framework that are congenial. This section will lay out fresh examples of how to insure inference validity or prove it is not possible. Such considerations are important for the statistical agency because it may wish to avoid SDL frameworks that are harmful to inference validity for high priority use cases.

6 Global confidentiality risk, multiple use cases, and social welfare maximizing decision-making for scientific research using confidential data

This section generalizes the framework in (Abowd *et al.*, 2019) to allow multiple use cases and multidimensional confidentiality risk assessment. Trade-offs in accuracy across multiple use cases, including the possibility of suppression of statistics, and trade-offs in the confidentiality risk across the multiple use cases will both be studied.

7 Recommendations for future research

This section will make suggestions for high value-added research in support of SDL frameworks that allow researchers to cooperate effectively with the statistical agencies that curate

data they wish to analyze. I will also comment on the challenge to users of data that incorporate both SDL and other quantifiable uncertainties as noted by boyd and Sarathy (2022) and Steed *et al.* (2022).

References

Abowd, John M., and Hawes, Michael B. 2023. Confidentiality Protection in the 2020 US Census of Population and Housing. *Annual Review of Statistics and Its Application*, **10**(1), null.

Abowd, John M., and Schmutte, Ian M. 2015a. Economic Analysis and Statistical Disclosure Limitation. *Brookings Papers on Economic Activity*, 221–267. `http://www.brookings.edu/~/media/Projects/BPEA/Spring-2015-Revised/AbowdText.pdf?la=en`.

Abowd, John M., and Schmutte, Ian M. 2015b. *Economic Analysis and Statistical Disclosure Limitation: Appendix*. `https://ecommons.cornell.edu/handle/1813/40581`.

Abowd, John M., and Schmutte, Ian M. 2019. An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. *American Economic Review*, **109**(1), 171–202. `https://doi.org/10.1257/aer.20170627`.

Abowd, John M., Schmutte, Ian M., Sexton, William N., and Vilhuber, Lars. 2019. Why the Economics Profession Must Actively Participate in the Privacy Protection Debate. *AEA Papers and Proceedings*, **109**(May), 397–402.

boyd, danah, and Sarathy, Jayshree. 2022. Differential Perspectives: Epistemic Disconnects Surrounding the U.S. Census Bureau's Use of Differential Privacy. *Harvard Data Science Review*, jun 24. `https://hdsr.mitpress.mit.edu/pub/3vj5j6i0`.

Cox, Lawrence. 1995. Network models for complementary cell suppression. *Journal of the American Statistical Association*, **90**(432), 1453–62.

Dinur, Irit, and Nissim, Kobbi. 2003. Revealing Information While Preserving Privacy. *Page 202–210 of: Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. PODS '03. New York, NY, USA: Association for Computing Machinery.

Duncan, George T., Keller-McNulty, Sallie A., , and Stokes, S. Lynne. 2001. *Disclosure Risk vs. Data Utility: The R-U Confidentiality Map*. Tech. rept. `https://www.niss.org/sites/default/files/technicalreports/tr121.pdf`.

Dwork, Cynthia. 2006. Differential privacy. *Pages 1–12 of: Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP)*.

Dwork, Cynthia, and Lei, Jing. 2009. Differential Privacy and Robust Statistics. *Page 371–380 of: Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*. STOC '09. New York, NY, USA: Association for Computing Machinery.

Dwork, Cynthia, and Naor, Moni. 2010. On the Difficulties of Disclosure Prevention in Statistical Databases or The Case for Differential Privacy. *Journal of Privacy and Confidentiality*, **2**(1).

Garfinkel, Simson, Abowd, John M., and Martindale, Christian. 2018. Understanding Database Reconstruction Attacks on Public Data: These Attacks on Statistical Databases Are No Longer a Theoretical Danger. *Queue*, **16**(5), 28–53.

Gong, Ruobin, and Meng, Xiao-Li. 2020. Congenial Differential Privacy under Mandated Disclosure. *Page 59–70 of: Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*. FODS '20. New York, NY, USA: Association for Computing Machinery.

Heffetz, Ori, and Ligett, Katrina. 2014. Privacy and Data-Based Research. *Journal of Economic Perspectives*, **28**(2), 75–98.

Hotz, V. Joseph, Bollinger, Christopher R., Komarova, Tatiana, Manski, Charles F., Moffitt, Robert A., Nekipelov, Denis, Sojourner, Aaron, and Spencer, Bruce D. 2022. Balancing data privacy and usability in the federal statistical system. *Proceedings of the National Academy of Sciences*, **119**(31), e2104906119.

Jarmin, Ron S., Abowd, John M., Ashmead, Robert, Cumings-Menon, Ryan, Goldschlag, Nathan, Hawes, Michael B., Keller, Sallie, Kifer, Daniel, Leclerc, Philip, Reiter, Jerome P., Rodríguez, Rolando A., Schmutte, Ian, Velkoff, Victoria A., and Zhuravlev, Pavel. *An In-Depth Examination of Requirements for Disclosure Risk Assessment.* under review.

Kifer, Daniel, Abowd, John M., Ashmead, Robert, Cumings-Menon, Ryan, Leclerc, Philip, Machanavajjhala, Ashwin, Sexton, William, and Zhuravlev, Pavel. 2022. *Bayesian and Frequentist Semantics for Common Variations of Differential Privacy: Applications to the 2020 Census*. `https://arxiv.org/abs/2209.03310`.

Slavković, Aleksandra, and Seeman, Jeremy. 2023. Statistical Data Privacy: A Song of Privacy and Utility. *Annual Review of Statistics and Its Application*, **10**(1), null.

Steed, Ryan, Liu, Terrance, Wu, Zhiwei Steven, and Acquisti, Alessandro. 2022. Policy impacts of statistical uncertainty and privacy. *Science*, **377**(6609), 928–931.

Wasserman, Larry, and Zhou, Shuheng. 2010. A Statistical Framework for Differential Privacy. *Journal of the American Statistical Association*, **105**(489), 375–389.