PNAS COMMENTARY

Database reconstruction does compromise confidentiality

Sallie Ann Keller[a,b,1] and John M. Abowd[a,1,2]

For decades, national statistics offices have struggled to answer a seemingly simple question, one for which Dick et al. (1) provide a compelling answer and "sober warnings:" What is the disclosure risk associated with publishing large numbers of aggregate statistics derived from a common source? Recent technological advances have brought this question into sharper focus with profound consequences for how statistical organizations approach the process of mitigating disclosure risk in their official releases.

What Is Statistical Disclosure Limitation?

Statistical disclosure limitation (SDL) is the process of treating confidential data to protect the identity and responses of data subjects' information in the published data. The most effective SDL mechanisms are those that permit high-confidence statistical inferences about a population while minimizing the likelihood that the inclusion of any particular individual in the underlying data could be leveraged to inform higher-confidence inferences about that individual. Put differently, SDL is an exercise in promoting uncertainty about individual data subjects' records while maximizing confidence in the aggregate statistics computed from those records.

The US Census Bureau has long been a world leader in the research, design, and implementation of SDL safeguards for the protection of data subject confidentiality in published products (2). Historically, the Census Bureau's disclosure risk assessments, like those of most other statistical agencies, focused on a narrow set of possible attacks, including isolating unique data subjects in individual (or small sets of) data tables, and record linkage-based reidentification attacks on microdata products. For tabular data, such as demographic data, these risk assessments often considered "plausible deniability" about the accuracy of any resulting reidentification or inference about a data subject to be sufficient to protect confidentiality. On the other hand, for microdata releases, more stringent SDL methods were applied to protect against the perceived greater threat of direct record linkage-based reidentification attacks (3, 4).

What Is Database Reconstruction?

The proof of the database reconstruction theorem in 2003 (5) (now usually called the fundamental law of information recovery) demonstrated the problem of relying on plausible deniability in a single table. Dinur and Nissim proved that every time you release any statistic calculated from a confidential data source, you reveal, or leak, a small amount of private or confidential information. If you release too many aggregate statistics too accurately, the entire underlying confidential microdata source can be exposed for the variables and coding used in the published statistics. Some critics of the applicability of this result to SDL have argued that "too many statistics" means enough to completely reconstruct every microdata record in the confidential data. This is an unrealistic all-or-nothing standard that ignores the more salient implication of Dinur and Nissim's argument: that each and every statistic you publish can be used to reduce uncertainty and improve an attacker's inference about the data subjects reflected in those statistics. An attacker need not reconstruct the entire underlying database in its original form for data subject confidentiality to be at risk. An attacker could instead reconstruct a portion of those records with full or nearly full confidence, as Dick et al. argue. The attacker's confidence in the likelihood that a particular record has been accurately reconstructed is improved by examining alternative solutions. It is in exploring this latter type of attack that Dick et al. demonstrate, again, how real and worrisome reconstruction attacks can be. Although reconstruction attacks may not always be able to reconstruct an entire database in its original form, Dick et al. show how to obtain a confidence-based ranking of rows that could align with the input database, providing a guided map for an attacker to follow.

Dick et al.'s "Sober Warnings" for Statistical Agencies Merit Action

In their article, Dick et al. demonstrate the inherent vulnerability of statistical agencies' traditional approaches to protecting aggregate statistics, which assessed the disclosure risk associated with tabular data products with different techniques than those used for microdata releases (6, see chapters 4 and 5). Light-touch SDL methods designed to protect confidentiality through plausible deniability result in aggregate statistics of such precision that they are, essentially, insufficiently protected microdata.

Author affiliations: [a]Office of the Associate Director for Research and Methodology, US Census Bureau, Washington D.C. 20233; and [b]Biocomplexity Institute, University of Virginia, Charlottesville, VA 22904

Author contributions: S.A.K. and J.M.A. performed research; and wrote the paper.

Competing interest statement: The authors are both employees of the US Census Bureau and wrote this commentary as part of their government work. The authors have made numerous public appearances where they discussed the issues raised in this commentary, all as part of their government duties.

See companion article, "Confidence-Ranked Reconstruction of Census Microdata from Published Statistics," 10.1073/pnas.2218605120

[1]S.A.K. and J.M.A. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: john.maron.abowd@census.gov.

Published March 15, 2023.

The fact that Dick et al. use Census Bureau data releases for their work is not surprising because these are some of the most granular and detailed official statistics published from confidential sources. Dick et al. show that large-scale, nonconvex optimization techniques can exploit this vulnerability and "exfiltrate entire rows of sensitive data with confidence." Their empirical findings echo and further confirm vulnerabilities that the Census Bureau's own research has demonstrated (7) and provide new, powerful tools for potential attackers. Their warnings have not fallen on deaf ears within the federal statistical community.*

> **"An attacker need not reconstruct the entire underlying database in its internal schema for data subject confidentiality to be at risk."**

How Does a Statistical Agency Move Forward?

The technological advances that now permit database reconstruction at scale have rendered the traditional approaches to SDL for aggregate statistics discussed above obsolete. It is only a matter of time before a malicious actor replicates the types of reconstruction attacks that Dick et al. and the Census Bureau have demonstrated. So, what should a statistical agency do? In this, there are lessons to be taken from the domain of cybersecurity. When an organization discovers a zero-day vulnerability in their software or a hole in their firewall, it would be unethical and irresponsible for them to wait for a hacker to exploit the vulnerability before taking corrective action. Statistical agencies should likewise strengthen their SDL methods to address the threat of database reconstruction now if they are to maintain the trust of their data providers and of the public at large. To this end, Dick et al. offer some advice. They note that, "[t]he only defenses against [reconstruction attacks] are to introduce imprecision in the underlying statistics themselves, as techniques like differential privacy do." This was precisely what the Census Bureau elected to do in the context of the 2020 Census (8). Leveraging more than a decade of research on formal privacy and working closely with data users to optimize the inherent privacy/utility trade-off, the Census Bureau was able to reduce the vulnerability of census tabulations to reconstruction while yielding data of higher accuracy than could be achieved using legacy methods at comparable levels of protection [(7), Tables 4 and 5].

Looking ahead, increased adoption of formal privacy approaches to data protection can yield substantial benefits but may also prompt some challenging public policy discussions. For the privacy community, formal privacy's quantification of disclosure risk will enable more principled evaluations of agencies' disclosure review and mitigation strategies than were possible under legacy SDL frameworks, but this quantification will likely also lead to greater debate about what constitutes "sufficient" privacy protection and what the proper balance between privacy and societal utility for official statistics should ultimately be. For data users accustomed to the largely untouched precision of aggregate statistics protected using legacy methods, Dick et al.'s imperative to introduce imprecision into the totality of those statistics will not be welcome news. Formal privacy, however, can help in responding to data users' fitness-for-use concerns in a principled and structured way. By quantifying the potential disclosure risk of each statistic to be published, formal privacy allows agencies (ideally in consultation with their data users) the opportunity to distribute statistical noise across those statistics in such a way that overall utility and privacy can both be preserved. For large data products with diverse uses, however, this flexibility may lead to difficult public debates over the relative importance of more precise data for certain use cases over others.

Also, and critically important, the privacy-loss accounting of formal privacy mechanisms allows for comprehensive tracking of the incremental disclosure risk across each of many data products derived from the same underlying data (a feature known in the literature as "composition"). Formal privacy's flexibility and composition will also be important as these solutions are considered for the sample-based surveys and other data collections over the coming years. Statistical agencies collect and disseminate data as their primary social product. It is reasonable to expect those data to be used for the public good as much as possible. However, there is no denying the increased risks demonstrated by Dick et al. While we wait for research that improves the quantification of specific incremental disclosure risks, we should also expand research on more targeted public-use products and more tiered access for research projects.

The threat of database reconstruction for tables from sample-based surveys, though perhaps less intuitive than for population-universe-based data tables at first glance, demands serious consideration given the prevalence of population uniques, as noted by Dick et al. and studied by Rocher et al. (9), because the more questions tabulated, the greater the chances of uniqueness. With over 44% of the US population having a unique combination of sex and age at the census block level, the likely overlap between sample uniques in reconstructed survey data (particularly at lower levels of geographic aggregation) and uniques in the broader population, and the corresponding implications for disclosure, should not be discounted. We have confidence that formal privacy research in the coming years will help to mitigate these risks.

ACKNOWLEDGMENTS. We are grateful to Michael Hawes for assistance in preparing this manuscript. The views expressed in this Commentary are those of the authors not the U.S. Census Bureau.

*In addition to substantial effort by individual statistical agencies, there have been numerous cross-government initiatives over the past several years to strengthen SDL protections for federal statistics. For example, the federal Commission on Evidence-based Policymaking (https://docs.house.gov/meetings/GO/GO00/20170926/106401/HHRG-115-GO00-20170926-SD001.pdf) and the federal Advisory Committee on Data for Evidence Building (https://www.bea.gov/system/files/2021-10/acdeb-year-1-report.pdf) have both encouraged greater investment and coordination across federal agencies for privacy-enhancing technologies, and the Federal Committee on Statistical Methodology has created a dedicated Subcommittee on Updating Statistical Methods for Safeguarding Protected Data (https://www.fcsm.gov/resources/safe-guard-data/), which is developing a Data Protection Toolkit to assist federal agencies in improving their SDL methods (https://nces.ed.gov/fcsm/dpt).

1. T. Dick *et al*., Confidence-ranked reconstruction of census microdata from published statistics. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218605120 (2023).
2. United States Census Bureau. A History of Census Privacy Protections. https://www.census.gov/library/visualizations/2019/comm/history-privacy-protection.html. Accessed 26 January 2023.
3. L. McKenna, Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing (Tech. Rep., U.S. Census Bureau, Washington, DC, 2018). https://www2.census.gov/ces/wp/2018/CES-WP-18-47.pdf. Accessed 1 March 2023.
4. L. McKenna, Disclosure Avoidance Techniques Used for the 1960 Through 2010 Decennial Censuses of Population and Housing Public Use Microdata Samples (Tech. Rep., Census Bureau, Washington, DC, 2019). https://www2.census.gov/adrm/CED/Papers/CY19/2019-04-McKenna-Six%20Decennial%20Censuses.pdf. Accessed 1 March 2023.
5. I. Dinur, K. Nissim, Revealing Information While Preserving Privacy PODS'03 (Association for Computing Machinery, New York, NY, 2003), pp. 202–210.
6. G. T. Duncan, M. Elliot, J. J. Salazar-Gonzalez, *Statistical Confidentiality: Principles and Practice* (Springer, New York, NY, 2011).
7. J. M. Abowd, M. B. Hawes, Confidentiality Protection, in the, US Census of population and housing. *Ann. Rev. Stat. Appl.* **10**, 2023 (2020).
8. J. M. Abowd, *et al*. The 2020 Census disclosure avoidance system topdown algorithm. Harvard Data Sci. Rev. (Special Issue 2) (2022).
9. L. Rocher, J. M. Hendrickx, Y.-A. de Montjoye, Estimating the success of re-identifications in incomplete datasets using generative models. *Nat. Commun.* **10**, 3069 (2019).