An In-Depth Examination of Requirements for Disclosure Risk Assessment

by

Ron S. Jarmin
U.S. Census Bureau

John M. Abowd
Cornell University

Robert Ashmead
U.S. Census Bureau

Ryan Cumings-Menon
U.S. Census Bureau

Nathan Goldschlag
U.S. Census Bureau

Michael B. Hawes
U.S. Census Bureau

Sallie Ann Keller
U.S. Census Bureau and U. of Virginia

Daniel Kifer
U.S. Census Bureau and Penn State

Philip Leclerc
U.S. Census Bureau

Jerome P. Reiter
U.S. Census Bureau and U. of Georgia

Rolando A. Rodríguez
U.S. Census Bureau

Ian Schmutte
Duke University

Victoria A. Velkoff
U.S. Census Bureau

Pavel Zhuravlev
U.S. Census Bureau

CES 23-49 **October 2023**

Abstract

The use of formal privacy to protect the confidentiality of responses in the 2020 Decennial Census of Population and Housing has triggered renewed interest and debate over how to measure the disclosure risks and societal benefits of the published data products. Following long-established precedent in economics and statistics, we argue that any proposal for quantifying disclosure risk should be based on pre-specified, objective criteria. Such criteria should be used to compare methodologies to identify those with the most desirable properties. We illustrate this approach, using simple desiderata, to evaluate the absolute disclosure risk framework, the counterfactual framework underlying differential privacy, and prior-to posterior comparisons. We conclude that satisfying all the desiderata is impossible, but counterfactual comparisons satisfy the most while absolute disclosure risk satisfies the fewest. Furthermore, we explain that many of the criticisms levied against differential privacy would be levied against any technology that is not equivalent to direct, unrestricted access to confidential data. Thus, more research is needed, but in the near-term, the counterfactual approach appears best-suited for privacy-utility analysis.

Keyword*:* federal statistical system, data disclosure risk, data access

---

[*] The views expressed in this perspective are those of the authors and not the U.S. Census Bureau. A slightly shorter version of this article with more of the material in the main text moved to the appendices is forthcoming as an invited Proceedings of the National Academy of Sciences Perspective: (https://doi.org/10.1073/pnas.2220558120). We thank the editor, Ariel Pakes, and referees from that journal for their careful review and comments.

Introduction

Data products from statistical agencies are vital to the functioning of many countries—supporting the distribution of political representation, business decision-making, research, government planning, and policy-making. The social welfare derived from these data products depends on their utility and also their confidentiality protections, which are mandated by legal and ethical considerations. By law, these confidentiality protections generally extend to all respondent information, regardless of how sensitive any particular response element may appear to be [12]. While statistical agencies regularly enhance confidentiality protections for obviously sensitive elements, the disclosure of which would likely cause harm to a respondent (for example, sensitive medical information), in practice statistical agencies' reliance on the voluntary participation of respondents and the information that those particular respondents may or may not consider sensitive is inherently contextual in nature [76]. This means that it would be impractical for agencies to discount the importance of confidentiality even for response information that may not readily be seen as particularly sensitive.

Since 1978, the federal statistical system has acknowledged that any public release of data generated from a confidential data source carries some risk of disclosure [39]. Balancing data utility and disclosure risk is often accomplished using a disclosure avoidance system, whose input consists of the confidential data and whose output consists of data products with a certain degree of confidentiality protection.

A disclosure avoidance system can apply a variety of statistical disclosure limitation (SDL) techniques to the input data, including suppression, coarsening, noise infusion, and aggregation (see Appendix B for a comparative discussion). The design of these systems and the tuning of their various parameters (e.g., underlying population thresholds or the amount of noise to be added) requires a careful cost-benefit analysis. The benefit side depends on stake-holder engagement and consideration of a wide range of use-cases. Quantifying the cost side requires a disclosure risk assessment methodology. This is a tricky task that is difficult to perform well—many notions of disclosure risk that at first seem intuitive can leave confidential data vulnerable by underestimating risk, or damage the data utility by overestimating risk. Quantifying disclosure risk is not a solved problem, but significant progress has been made in the last few decades, partly driven by systematic approaches that avoid presupposing a specific measure of disclosure risk. Instead, they propose a set of desirable properties and then identify specific disclosure risk measures that satisfy those properties.

In this paper, we illustrate such an approach in a minimally technical way. We lay out a set of relatively non-controversial desiderata for disclosure risk assessment methods. We use these desiderata to analyze three major frameworks: *absolute disclosure risk*, *prior-to-posterior comparisons*, and the *counterfactual comparisons* that motivate differential privacy [34, 32, 85]. Based on this analysis, we conclude that it is impossible to simultaneously satisfy all the desiderata. However, differential privacy and its counterfactual framework are supported by the best available science, clearly perform the best and, therefore, should be recommended for use at statistical agencies when SDL implementations based on this framework can be effectively implemented. The potential for effective implementation is an important caveat here, because of the inherent technical challenges associated with implementing noise infusion-based SDL solutions within a differentially private risk assessment framework, especially for complex weighted survey designs. Similarly, any particular SDL implementation (including those within a differential privacy risk assessment framework) may be better or worse at maximizing utility, depending on that implementation's design and the parameter choices reflected therein. Thus, the adoption of differential privacy as a risk assessment framework should be seen as entirely distinct from the selection of any specific SDL implementation (good or bad) informed by that risk assessment. Other risk assessment approaches also offer value,

but need more research to address important shortcomings.

In a recent paper, Hotz et al. [53] called for a moratorium on the use of differential privacy. They argued that it is incompatible with cost-benefit analyses of releasing new data, and that it prevents the computation of "an essentially unconstrained number of statistics" [53]. The first claim is incorrect—it is contradicted by real-world deployments of differential privacy (including the 2020 Census—see Section 1 and Appendix A). As we explain in this paper, the second claim could be levied against *any* method that controls disclosure risk (see, for example, criticisms of cell suppression in earlier censuses [69]). Consequently, the question of whether "an essentially unconstrained number of statistics" should be supported is not a matter of choosing between differential privacy or an alternative disclosure avoidance technology. Instead it is a question of whether to protect confidentiality at all.

The quantity, granularity, and precision of published statistics all contribute toward the overall disclosure risk of the data. To protect confidentiality for its 100+ recurring censuses and surveys, the Census Bureau makes use of a wide variety of disclosure avoidance methods, including various implementations of suppression, coarsening, and noise injection approaches, including but not limited to those based on differential privacy. Because the application of any disclosure avoidance method reduces the utility of the resulting data in non-trivial ways, and because the choice of any particular disclosure avoidance method and how it is implemented can have differing impacts on utility for diverse subsets of use cases, the Census Bureau takes great care to select the best available disclosure avoidance method for each particular census and survey it conducts and for each data product it releases. These choices, informed by external stakeholder feedback, help maximize the resulting utility of published data for those uses that provide the greatest societal value. However, they are made with the understanding that protecting confidentiality requires that utility for some other potential or unanticipated use cases may suffer as a result. While these trade-offs may not be apparent for smaller data products (those with fewer statistics or less granularity), or for those that are used for only a narrow range of use cases, they will be visible and inescapable for large-scale, highly-granular data products with many diverse use cases, such as the 2020 Census.

This paper is organized as follows. In Section 1, we briefly review activities performed by the Census Bureau to quantify the utility of 2020 Census data products. In Section 2, we list some generally accepted requirements that disclosure risk measures should satisfy. Although they are simple enough to be discussed in a non-mathematical way, they are still rich enough to provide meaningful insights. In Section 3, we overview three of the major methodologies for measuring disclosure risk. In Section 4, we use the desiderata to investigate the suitability of these disclosure risk assessment methodologies. We discuss implications for the federal statistical system in Section 5 and present conclusions in Section 6. The Appendices include: A. Summary of how the Census Bureau used stakeholder feedback to assess how the SDL applied to the 2020 Census affected accuracy; B. Comparison of different SDL methodologies that explains why the combination of aggregation and noise infusion was used for the 2020 census; C. Comparison of synthetic data methods with traditional SDL; D. Discussion of cost-benefit and social welfare analyses of disclosure risk; and E. Detailed analysis of published claims about the 2020 Census Disclosure Avoidance System and risk analysis.

1 Quantifying Utility via Stakeholder Engagement

Population censuses have many uses. It is unrealistic to quantify their social benefit with a single number. For the 2020 Census, the Census Bureau developed a suite of metrics, based on stakeholder engagement, that were used to understand the impact of different levels of accuracy on stakeholder

use-cases. These metrics guided the development and tuning of the 2020 Census Disclosure Avoidance System (DAS) and clearly document the extensive balancing of interests that occurred [7]. See the details in Appendix A.

2 Desiderata for Disclosure Risk Assessments

Suspected privacy breaches for Census Bureau data products, protected using traditional methods, are routinely reported to the Chief Privacy Officer or the Disclosure Review Board. They are investigated but, as a matter of policy, the agency generally does not publicly confirm that a specific suspected breach is an actual breach. The agency has released some breach summaries [71].

Many breaches also go unreported—what is the incentive for an attacker to reveal a breach that is being exploited? Furthermore, reporting confidentiality breaches is sometimes discouraged by curating organizations, e.g., IPUMS International ("[a]lleging that a person or household has been identified is also prohibited" [54]).

A typical argument used to justify weaker privacy protections is "there have been no breaches so far." Such statements are often demonstrably false [71] and also illustrate the belief that loss of trust due to privacy breaches is easy to overcome and should not affect data quality. There is no evidence to support this belief, and strong evidence against it [22].

As one prominent state demographer recently pointed out at a Federal State Cooperative for Population Estimates meeting "in past censuses, it was possible to create a PUMS-like [Public-Use Microdata Sample] file from the SF1 tables [i.e., decennial census publications]. It was just a pain in the neck" (see Appendix F). Whether the reconstruction of microdata records for possible demographic or redistricting uses should be considered a privacy breach is a question requiring a nuanced response. However, the knowledge that it is attempted even by those who accept the requirements of confidentiality protection illustrates that the strengths and limitations of different methods of protection must be taken seriously, and that defensible methodologies for disclosure risk assessment are essential.

We now discuss desiderata that should be considered when assessing methodologies that quantify disclosure risk. It is important to note that a disclosure risk assessment methodology is not the same thing as a disclosure limitation mechanism. A mechanism M is a piece of software whose input is confidential data $\mathcal{D}$ and whose output consists of data products R that are suitable for public release. The role of the assessment methodology is to measure the disclosure risk of the data products in a principled manner. Disclosure risk assessment methodologies are not mutually exclusive, meaning that privacy properties of M can be analyzed by multiple methodologies, though of course only methodologies that are defensible should be used for assessment. The aim here is to present reasonably uncontroversial desiderata for risk assessment methodologies. Even with these simple desiderata, this exercise provides insights about various disclosure risk assessment methodologies.

2.1 Transparency

The principle of transparency is that details (e.g., source code and algorithm parameters) of a mechanism M should be available to the public. The methodology for disclosure risk assessment should consider the possibility that an attacker would use the details of M to attempt to reverse-engineer [24] the public data products R to extract information about the confidential data $\mathcal{D}$ as, for example, can be done with many mechanisms supporting k-anonymity [84].

Transparency is necessary so that users of the public data products R can understand what errors, limitations, or systematic biases the data might have [25, B1]. This understanding facilitates

public privacy vs. utility discussions and makes it possible to quantify utility through stakeholder engagement, as explained in Appendix A. Transparency is also necessary for the public to verify that M was implemented correctly. A 2010 analysis by Davern et al. [10] revealed mistakes in the implementation of earlier SDL systems, with important implications for data quality. These mistakes were only detected because there were multiple sources of information available about the same population and estimates from the different sources were inconsistent. When multiple sources are unavailable, mistakes may not be detectable unless the code of M is public. Abowd and Schmutte [1] call this property of SDL "discoverable" and point out that unless the system is either discoverable or public there is a strong risk that the likelihood function of the released data (the probability distribution used by statisticians to make inferences) suffers from nonignorability (a systematic bias like nonignorable missing data [67]), making correct scientific inferences difficult.

Many uses of SDL are *not* transparent. In particular, the 2010 Census used non-transparent methods (i.e., data swapping with a confidential algorithm and parameters) due to concerns that knowledge of implementation details could be used to undo the privacy protections. This resulted in a general lack of awareness that data had been modified. There is little understanding of how non-transparent SDL affects analyses (in particular, small area estimation [8, e.g., Table 1]). In contrast, in the 2020 Census, the Census Bureau embraced transparency, publishing source code of the 2020 DAS [90], providing demonstration data products using historical data [96], developing and releasing associated summary metrics, and triggering a robust debate on the proper role, and balance, of privacy and utility in the federal statistical system [45].

2.2 Universality

A method for disclosure risk assessment should be applicable to any type of data product: microdata samples, tabulations, regression coefficients, etc., even if the agency only plans to release a single product because all data products contribute to disclosure risk, and their cumulative impact must be assessed when, inevitably, additional products are proposed from the same source. While some proposals call for the release of only microdata, restricting an assessment method to only work on microdata is too limiting. For example, researchers may wish to publish regression coefficients from a Federal Statistical Research Data Center (FSRDC). In general, the format of a data product should be decided as part of the privacy/utility cost-benefit analysis, and a disclosure risk assessment methodology should support this analysis by providing a privacy cost for any data product format.

2.3 Deem releasing uninformative statistics not a disclosure risk

Consider a data release that produces something completely uninformative, like a heavily perturbed total population count. This information should be considered a minimal disclosure risk. This is a sanity check but, as we shall see, some disclosure risk assessment methodologies fail it.

2.4 Support arbitrary statistical analyses

The types of questions that researchers and policymakers try to answer often involve complex data analyses. This raises an important concern—would a disclosure risk assessment methodology stand in the way of these activities? We split this issue into two desiderata: (1) a disclosure assessment method should allow the creation of data products that accurately support an arbitrary, *pre-specified* statistical analysis, and (2) the assessment method should allow the creation of data products that support arbitrary analyses that have *not* been specified in advance. We discuss the first desideratum

here and the second desideratum, which turns out to be equivalent to supporting arbitrarily many kinds of statistical analyses, in 2.5.

Some analyses are inherently incompatible with any reasonable conception of privacy—for example, analyses designed to determine the identities of a specific person's sexual partners or parties to payments for illicit goods. A public health authority or law enforcement agency, respectively, might have legitimate reasons to undertake such investigations, but such activities are clearly prohibited for a U.S. statistical agency by the definition in Statistical Policy Directive 1 (SPD 1, 44 U.S.C. § 3561(8)) of "nonstatistical purpose." Any pre-specified *statistical* analysis–for example, aggregate counts, regression designs, learning models—for a *large enough* population should be supportable.

This desideratum includes tiered access and restricted data use environments like the FSRDCs, meaning an analysis need not be pre-specified at the time of the creation of a data product. Instead, researchers can analyze public data products and when preliminary results indicate potentially useful findings (or that a potentially useful investigation cannot be supported by existing data products), the researchers can specify the analyses they need to perform and use the FSRDCs to conduct them. The results of the work in the FSRDC can then be protected with privacy technology and publicly released, as Chetty et al. did with IRS and Census Bureau data [20] and with Facebook data [19]. Note, however, that quantifying privacy loss due to exploratory analysis of raw data, known as the "Fienberg Problem" [38], is an open research question.

2.5 Support arbitrarily many analyses

Given the explosive growth of various research communities with different needs, goals, and statistical methodologies, stakeholders may wish to avoid FSRDCs and would be delighted to receive data products that support arbitrary analyses that are not specified in advance. Certainly, this would be less cumbersome for them. Since the analyses are not known in advance, the only way to support them at the time of data product creation is to ensure the data products can support arbitrarily *many* analyses. However, providing a data product that accurately supports arbitrarily many analyses is mathematically equivalent to releasing confidential data $\mathcal{D}$ with no protections at all [27]—in other words, the confidential data could be reconstructed from those data products. Therefore neither differential privacy nor any other methodology for effectively assessing disclosure risk can legitimately satisfy the criterion of accurately supporting arbitrarily *many* analyses for any non-trivial degree of confidentiality protection – it is a mathematical impossibility.

2.6 Deem nearly perfect reconstruction of high-dimensional data or nearly perfect inference about a data subject high disclosure risk

This desideratum acknowledges that data products that are nearly equivalent to the original confidential microdata (i.e., if microdata can be accurately reconstructed from the released product) is a disclosure risk if the original microdata were considered confidential. Note that, as discussed earlier, this principle is in conflict with support for arbitrarily many analyses.

This is true even for data collected from sample surveys. In addition to enabling inference based on small homogeneous groups, near-perfect reconstruction renders sample uniques easily identified, and sample uniqueness is often a strong predictor of population uniqueness. The challenge of detecting and protecting a population unique—a record in the response data that represents the only entity in the population with the associated subset of characteristics—has a long history of study in surveys and censuses [16] as well as in biostatistics [52]. Prior work shows particularly strong results without requiring access to the confidential data for verification [78, 31].

Reconstruction of an entire dataset is not necessary for a prohibited disclosure to occur. For example, suppose a dataset reveals that all 10 people in zip code 12345 who are 43-year-old Asian Hispanic males have cancer. The confidentiality of these individuals has been breached even if the enumeration was impossible.

2.7 Deem generalizable inference not a violation of confidentiality

Dwork and Pottenger [30] proposed explicitly acknowledging that generalizable statistical inference is not a privacy breach because it does not jeopardize the confidentiality of the underlying data. Data products are published to support generalizable inferences. Statistical inferences from the data can still be harmful, but the harm from such inferences should not be assigned to the SDL system. In the case of sample surveys, for example, one hopes that the findings from those data generalize to the entire population. In the case of censuses, one hopes that the results generalize to subsequent years. For example, a policy implemented on the basis of the 2020 Census should still succeed in achieving its primary aims as the population grows and changes.

Thus, one should expect to learn something new from the data, and this knowledge should not be considered a breach of confidentiality. A canonical example [30] is the large scale cancer study CPS-I spanning the 50s, 60s, and 70s [11, 49], which conclusively established the link between smoking and cancer. This finding means that insurance rates for smokers may have increased after the study was published, which would be an undesirable outcome from their point of view. However, if a smoker was not even a study participant, should this improved inference about their cancer risk be considered a privacy breach? According to this desideratum, it would not, since this inference was clearly based purely on generalizable knowledge.

Now suppose a different smoker did participate in the study. Inferences about that person's cancer risk would result from a combination of generalizable knowledge obtained from the data and use of the subject's data in the production of the report. By accepting this desideratum, one is expecting a disclosure risk assessment methodology to be able to determine how much of an inference is due to generalizable knowledge and how much is due to participation in the dataset on which the inference is based.

There is a separate policy concern regarding the publication of data on sensitive subjects or populations. For example, if the Census Bureau learned that the main users of one of its data products were foreign intelligence agencies trying to manipulate U.S. elections, it would probably stop releasing this data product on policy grounds, even if the foreign intelligence agencies were only doing generalizable statistical analysis and no prohibited disclosure occurred. That is correctly a policy decision independent of the concerns about confidentiality but consistent with the guidelines in SPD 1 (44 U.S.C. § 3561-4).

Rejecting this desideratum necessarily conflates decisions on appropriate statistical uses with those concerning disclosure limitation and risks harming the scientific value of public data sets.

2.8 Support composition

Imagine a scenario in which one set of data products was deemed a low privacy risk, another set of data products was also deemed low privacy risk, but their joint release was deemed to be a catastrophic privacy violation. The technical terminology is that the two privacy-protected data products did not *compose* well–the cumulative effect of all publications from the same confidential data source(s) contributed to the disclosure risk in ways that could not have been predicted from separate, independent analyses [42]. Separate "release and forget" analyses have been standard practice for SDL practitioners in the past, as was the case for disclosure risk reviews done on data

products from the 2010 Census [69, 70], but they provide no indication of the cumulative disclosure risk of all publications, unless the disclosure assessment methodology supports composition.

A disclosure risk assessment method composes well when an upper bound on the joint privacy risk for a set of data products is predictable based on a privacy analysis conducted on each data product separately. This property simplifies the design of disclosure limitation mechanisms and allows different parts of an organization to plan different product releases with minimal coordination as long as there is centralized review and control of the cumulative disclosure risk. It also simplifies the marginal cost-benefit analysis of each new data publication. For instance, one could separately analyze the incremental privacy cost of each new data publication and weigh it against its incremental utility.

Rejecting this desideratum means organizations must plan their data releases sequentially, measuring the cumulative effect of all prior releases from the same confidential source, as well as the proposed new data release. This will affect the timeliness of data product releases because an organization would release one data product, then must conduct an analysis of what happens when a second data product is added. Subsequently, when planning a third data product, it must be jointly analyzed with the first two. Adding each new data product becomes an increasingly complex endeavor. This is a limitation that, for example, currently impacts all statistical products that use properly implemented primary/complementary suppression, as discussed in [8].

2.9 Acknowledge multiple types of attackers and auxiliary information

When publicly releasing data products, privacy risks can come from a variety of sources. For example, information technology giants may combine the released data with their own user data to obtain information they did not previously possess, as their data can differ in quality, extent, and measured attributes when compared to carefully worded survey data. Another source of privacy risk comes from data brokers who buy and sell data about individuals in bulk, companies that engage other companies to access their customers and data, prying neighbors, redistricters who want personal data as they can to predict individual voting patterns, or government agencies that are enjoined by law from requesting or issuing subpoenas for the confidential data.

The acknowledgment and analysis of privacy risks due to multiple different attackers (who do not necessarily share data) is critical to avoid mistakes. For instance, if the only type of attacker considered is an equivalent of Amazon (e.g., has names, addresses, and purchasing histories of a large portion of the population), then a disclosure limitation methodology based solely on that attacker could lead to the erroneous conclusion that publishing addresses and consumption patterns of survey participants does not introduce any risks since the data are "already out there." Other attackers may not have Amazon's knowledge, and SDL should prevent them from learning personal information they don't already have even though it is "out there."

2.10 Resist brittleness

A disclosure risk assessment exhibits brittleness if the inclusion or removal of data about one entity materially compromises the privacy of another entity's data. Confidentiality protection for businesses can be brittle in this sense because whether or not a competitor supplied data can affect the confidentiality protection afforded to other businesses in the same market. An agency does not exclusively control information about any particular data subject—the data subject can do whatever they want with their own information. Thus a disclosure risk assessment method should anticipate such a possibility and should not be brittle—whether or not a data subject decides to release their own data publicly or to the agency should not materially affect the assessed privacy

risks of everyone else.

2.11 Be resilient to misspecification

Many disclosure risk assessment methodologies make assumptions about capabilities of the attackers—the data they may or may not possess, the probability distribution that encapsulates their prior beliefs, the attack method they may choose to use, and even which pieces of information should be considered sensitive. What if the assumptions are wrong or obsolescent (e.g., when new external datasets become available in the future)? This is similar to model misspecification. Disclosure risk assessment methods should be able to quantify how they are affected by deviations from their underlying assumptions.

2.12 Control computational effort

As we saw with composition, computational effort can become a major practical consideration when creating data products. Computational expenses can be incurred when including all prior data products in a disclosure risk analysis, when computing posterior probabilities in Bayesian methods, and generally when handling large data sets. Ideally, reducing computational expense should not come at the cost of weakening the disclosure risk assessment (e.g., making assumptions about prior distributions that are clearly incorrect, but allow for closed-form estimates).

2.13 Control manual effort

The development of SDL systems typically requires expertise in statistical modeling, computational algorithms, and optimization. The use of a disclosure risk assessment methodology incurs an initial cost in staff training. It may also incur marginal costs in keeping track of data that are available elsewhere (and that may be used by an attacker) and analyzing disclosure risks of new data products. It should also not hinge on the assumption that personnel are more adept at finding viable attack vectors than actual potential attackers. Keeping the manual effort manageable is another important goal for any statistical agency.

3 Major Disclosure Risk Assessment Methodologies

Having identified several major desiderata for evaluating prospective disclosure risk assessment methodologies, we next consider some of the major methodologies that have been proposed for this risk assessment: absolute disclosure risk, prior-to-posterior comparisons (a type of *relative* disclosure risk), and counterfactual posterior-to-posterior comparisons (another type of *relative* disclosure risk). To keep the mathematical detail at a minimum, only the motivations and broad outlines are described, but this is still enough to compare the different frameworks along the desiderata outlined in Section 2.

3.1 Absolute disclosure risk

Absolute disclosure risk can be formulated in different ways. For concreteness, we follow a specific proposal from Hotz et al. [53] in the following. The confidential data set $\mathcal{D}$ consists of records of n data subjects $p_1, \ldots, p_n$. Each person p_j has sensitive information s_j. An attacker has some data A (e.g., an external dataset such as voter registration records, or personal observation) and a Bayesian prior, $P(s_j = S \text{ and record } r_i \text{ belongs to person } p_j)$. Since this risk assessment approach requires linkage to external information, it also requires that the data product R produced by

mechanism M be either a dataset of records $\{r_1, \ldots, r_n\}$ (e.g., a public microdata file), or that a partial or complete microdata file could be reconstructed from non-microdata data products (e.g., reconstruction of a dataset of records $\{r_1, \ldots, r_n\}$ from tabular data releases).

To attack a person p_j, the attacker must first determine which protected record r_i corresponds to that person and then must infer a value S for s_j, the sensitive information about p_j. For discrete attributes, the probability of success is written as $P(s_j = S \text{ and } r_i \text{ belongs to person } p_j \mid A, R, M)$, where conditioning on M indicates that the mechanism is known but the realized values of its random variables, if any, are not known. Computing this probability requires making assumptions about the attacker's prior knowledge A and prior beliefs about which distribution(s) could have generated the data. We refer to this probability as the **absolute disclosure risk with linking**.

Hotz et al. [53] present an analogy to pathology, in which this absolute disclosure risk is analogous to the probability that someone will die within x months—clearly a number that people would care about. We extend this analogy when discussing prior-to-posterior and counterfactual comparisons as it presents a vivid way of highlighting the differences in methodologies.

We also consider a modification of this proposal that we refer to as **absolute disclosure risk without linking**. In this setting, an attacker only needs to learn the sensitive information s_j about person p_j without having to specifically identify which privacy protected record r_i belongs to p_j. As demonstrated by the attack of Homer et al. [52] on genomics data, linking is not necessary for serious disclosures to occur because the probability of correct attribute inference can be accurately estimated from other properties of the data [37]. Dropping the linking requirement also adds more flexibility since a mechanism M no longer needs to produce a record-level dataset as the public data product R—it can produce regression coefficients, tabulations, a deep learning model, etc. The success probability of the attacker can then be written as $P(s_j = S \text{ and person } p_j \text{ is in } \mathcal{D} \mid A, R, M)$.

3.2 Prior-to-posterior comparisons

Continuing the analogy to pathology, while absolute disclosure risk corresponds to the overall probability of dying within x months, prior-to-posterior comparisons are comparisons between the current probability of dying and the probability of dying after a treatment or other activity, thus trying to isolate the effect of the treatment. Here, the treatment is the use of M to release privacy-protected data products.

In SDL applications, the attacker starts with a prior probability that the sensitive attribute s_j of person p_j equals some value S. Mathematically, the prior is $P(s_j = S \text{ and person } p_j \text{ is in } \mathcal{D} \mid A)$. After the confidential data $\mathcal{D}$ are run through M to produce the public data products, the attacker can form the posterior distribution: $P(s_j = S \text{ and person } p_j \text{ is in } \mathcal{D} \mid A, R, M)$. These two probabilities can be compared, typically, but not always, by subtraction: posterior $-$ prior, or division: $\frac{\text{posterior}}{\text{prior}}$ to measure the change in beliefs induced by the public data products. Because such disclosure risk methods are comparisons (i.e., before and after R is released), they are known as measures of *relative disclosure risk*.

3.3 Counterfactual posterior-to-posterior comparisons

This framework is used by differential privacy and its variants [13]. Consider a clinical trial where the treatment group receives medicine for six months, and a control group receives a placebo for six months. Comparing the mortality of these two groups is analogous to counterfactual posterior-to-posterior comparisons [79]. In terms of disclosure limitation, this methodology considers two "worlds"—the *actual* world in which the target person's data are included in the dataset $\mathcal{D}$ that is sent to M, and the *counterfactual* world in which that person's data are omitted or replaced

by a blank record before being sent to M. The idea behind the counterfactual world is that a person may have a set of actions that can be performed unilaterally to protect privacy, such as not participating in a survey or submitting a blank response. If an individual considers these actions sufficient to protect their privacy, then the counterfactual world is a privacy-preserving baseline for that person.

The actual world and counterfactual world can be compared in several ways. The first, often called the *frequentist approach*, is developed from the perspective of hypothesis testing: based on the data products R and the source code of M, could the attacker detect if the full dataset was the input to M or whether the target person's record was removed before running M (i.e., could the attacker distinguish between the actual and counterfactual worlds)?

The second way of comparing the actual and counterfactual worlds is often called the *Bayesian approach*. One would compare the inference that an attacker would have in the actual world: $P(s_j = S \mid A, R, M,$ target p_j is in the data) to the inference the attacker would have made in the counterfactual world: $P(s_j = S \mid A, R, M,,$ target p_j was *removed* from the data). If the two conditional probabilities are similar it means that the attacker is likely to draw nearly the same conclusion even if the target individual had never participated in the survey. The two conditional probabilities can be compared by dividing one by the other, but more sophisticated comparisons are often used instead. A detailed treatment can be found in [64].

3.4 Discussion

Recalling that a disclosure risk assessment methodology is not the same thing as a mechanism M, consider a mechanism that adds noise to the sufficient statistics of a model and samples records from the model to produce a synthetic dataset. The overall effectiveness of this mechanism at mitigating disclosure risk can be analyzed from the perspective of all three disclosure risk assessment methodologies discussed above. Thus, common SDL methods like suppression, coarsening, swapping, etc. [99], are not discussed here as risk assessment methodologies since they are frameworks for designing mechanisms which can then be analyzed by disclosure risk assessment methodologies. We also note that each methodology is an umbrella that may have many variations depending on the definition of sensitive information, the types of probability distributions used by the attacker, the knowledge the attacker can possess, and how probabilities can be compared to each other.

It is often claimed that privacy-loss measures obtained from the counterfactual methodology place a hard limit on the privacy loss—that is, after a certain point, there can be no further interaction with the underlying confidential data. On the contrary, the counterfactual method and *all* other methodologies provide a measure of disclosure risk. What an agency does with the risk measures is up to the agency. It can weigh the incremental privacy risk against the incremental gain in utility of subsequent data publications. Any claims that rigid, dogmatic restrictions are pre-requisites to the use of various disclosure risk assessment methodologies are specious—the only requirement is that an agency must make decisions that are defensible.

4 Analysis of Methodologies

We next compare the three different methodologies from Section 3 using the desiderata from Section 2. We go through the desiderata in the same order. See Table 1 in Section 4.14 for a summary. The purpose of these desiderata is to help ensure that a methodology is chosen based on sound principles, but it also helps identify promising areas of future privacy research.

4.1 Transparency

In principle, all three methodologies *could* support transparency, and so the main question is whether the methodologies have a viable proof of concept. There are two main challenges for a proof of concept. *Computational*—can the relevant posterior probabilities $P(\cdot \mid R, M, A)$ be computed accurately and efficiently enough for decision-making purposes, even when M is complex? *Model specification*—where do the attacker's probability distributions and information come from, since computing the posterior distribution requires specifying $P(\mathcal{D})$, the prior probability of the data. If the data curator must estimate $P(\mathcal{D})$ from the actual data, then the methodology could become non-transparent because the data curator would either have to reveal $P(\mathcal{D})$, so that the public could verify the disclosure risk assessment, or would have to hide it, resulting in degraded transparency. If the data curator reveals a $P(\mathcal{D})$ estimated from the data, then this could itself lead to disclosure.

The methodology based on counterfactual posterior-to-posterior comparisons has many proofs of concept, most notably the literature on differential privacy. A comprehensive survey on the different ways of measuring disclosure risk under this methodology can be found in [26]. There are many computationally-friendly tools for accurately approximating or upper-bounding the privacy loss in this framework [17, 101]. Interestingly, even though posterior distributions may be difficult to compute by themselves, one only needs to *compare* two posterior distributions (for the actual and counterfactual worlds) and this comparison can be performed efficiently. The other challenge, model specification, is typically resolved by providing protections for absolutely all priors. For decision-makers who are used to analyzing re-identification probabilities, the counterfactual contrast can be converted into significance level and statistical power for the null hypothesis of determining participation in the dataset (a pre-requisite to re-identification) versus the alternative of non-participation [98, 64]. Such an approach uses trade-offs that statistical agencies have long employed in designing surveys. Other approaches use a smaller set of priors [48, 26], but their strengths and weaknesses require further study. This is one promising direction of research for this methodology. Thus the counterfactual framework supports transparency through many proofs-of-concept.

The prior-to-posterior comparison methodology also has some transparent proofs-of-concept [65, 83] but they are more challenging to implement. Model specification is handled by providing guarantees for a large set of priors or for priors that are easy to work with (such as the Gaussian distribution). Aside from the "easy" priors, which may not faithfully represent the data, these methods also encounter computational obstacles when computing posterior distributions. Monte-Carlo and MCMC simulations can often be used for these purposes [68]. However, for large-scale data sets, MCMC approaches can be slow to converge, potentially leading to inaccurate estimates of disclosure risk.

The absolute disclosure risk methodology, however, currently lacks a proof-of-concept. When the public data products R take the form of microdata, absolute disclosure risk with linking is typically measured by performing linking attacks between the confidential data $\mathcal{D}$ and R [77]. Such empirical assessments, however, are frequently criticized for using simplistic, rather than the best available, linking algorithms and external data, and often significantly *underestimate* disclosure risk [78]. In addition, they reduce transparency since the results, upon which decisions are based, cannot be audited by the public because such audits require access to the confidential data, which could, however, be granted to an auditor within a restricted-access environment. Absolute disclosure risk *without* linking faces the two open challenges: specifying a prior distribution $P(\mathcal{D})$ that would not lead to underestimation of disclosure risk, and computing the posterior distribution efficiently and accurately in large-scale data. Thus, currently only the prior-to-posterior and counterfactual methodologies have transparent proofs-of-concept, the latter being more numerous.

4.2 Universality

The absolute disclosure risk without linking, prior-to-posterior, and counterfactual methodologies all place no restrictions on the public data products R. Only absolute disclosure risk *with* linking restricts the form of R, because it requires the released product to be microdata or of a form that could be reconstructed into microdata. Thus, it would not be applicable to many common alternate forms of data releases, such as regression coefficients, that might be released from an FSRDC.

4.3 Deem releasing uninformative statistics not a disclosure risk

Consider a data collection effort that records personal information about individuals such as race, citizenship, etc. Releasing a "data product" that consists *only* of a heavily perturbed national population count would be treated as nearly zero disclosure risk under the prior-to-posterior methodology—the posterior probability of learning sensitive information after releasing such data is nearly the same as the prior probability of the sensitive information due to the non-informative data product. Similarly, it has nearly zero disclosure risk under the counterfactual framework (see the explanation of per-attribute semantics in [64]).

Absolute disclosure risk with linking is not applicable in this setting, but absolute disclosure risk *without* linking could be applied. In some situations, the absolute disclosure risk method would assign a large disclosure risk to the national count. For example, a target individual's neighbor may be completely certain that the target participated in the census and may have a very good guess about the target's race, marital status (e.g., same-sex or opposite-sex), etc. In some cases, even citizenship status would be entirely predictable for certain people. In other words, the prior belief $P(s_j = S \text{ and person } p_j \text{ is in } \mathcal{D} \mid A)$ is already large. The release of uninformative statistics would not change the beliefs and so the posterior $P(s_j = S \text{ and person } p_j \text{ is in } \mathcal{D} \mid A, R, M)$, which is the only quantity this methodology considers, would also be large. Therefore, the heavily perturbed count would not be releasable *even though the neighbor learned virtually nothing new about the target individual*. See also the Montana example in Section 4.7.

This desideratum suggests that a posterior probability needs to be compared to *something* in order to determine whether a data release changed an attacker's beliefs or reduced their uncertainty. Absolute disclosure risk specifically avoids such comparisons.

4.4 Support arbitrary statistical analyses

Tiered access can minimize the need for pre-specifying all potential uses of data in advance. Analyses that are formulated later, and which are not supported by public data products R, can then be performed in restricted access services like FSRDCs with the outcomes subsequently protected by SDL. While this can affect the pace of research, it currently appears to be unavoidable. In principle, absolute disclosure risk without linking, prior-to-posterior comparisons, and counterfactual comparisons can be used to analyze the privacy risks of the outcomes produced in restricted-access services, but absolute disclosure risk with linking may not be applicable.

When the discussion tilts towards known mechanisms M for different kinds of analyses, and whose privacy risks can be efficiently assessed, the counterfactual methodology has the most support and underlying research at the moment. Nevertheless, supporting a wider variety of analyses is an important research direction.

4.5 Support arbitrarily many analyses

The difference with the previous desideratum is the quantity and variety of arbitrary analyses. As mentioned in Section 24.5, the only way to allow researchers to conduct arbitrarily many analyses accurately on a public data product R is to violate confidentiality and simply make R identical to $\mathcal{D}$. This mathematical fact [27] must be acknowledged. No disclosure control methodology can achieve this goal.

4.6 Deem nearly perfect reconstruction of high-dimensional data or nearly perfect inference about an individual high disclosure risk

When the confidential data $\mathcal{D}$ can be nearly perfectly reconstructed from the public data R, absolute disclosure risk without linking, prior-to-posterior, and counterfactual comparisons would assess R as a high disclosure risk. When R takes the form of microdata, then in certain situations, absolute disclosure risk *with* linking will not characterize this release as high risk. For example, suppose the target individual is known to be in the dataset (or the dataset may be a full census). If the target individual shares the same linking attributes with 9 other people, and if all of their sensitive information is the same, then an attacker has learned the target's sensitive data without knowing which of the 10 records belongs to the target. This sensitive attribute inference is probably not a generalizable inference because it is based on a sub-population of only 10. It is the type of inference that can raise a concern about confidentiality, as noted for k-anonymity [84]. In a sense, the inability to link seems irrelevant, but absolute disclosure risk with linking discounts the perfect inference specifically because of the limitations on linkage, ignoring the success of the non-generalizable inference. This is another example of a simple situation that is not handled correctly by absolute risk assessment with linking, casting doubt on its proper behavior in more complex situations.

4.7 Deem generalizable inference not a violation of confidentiality

In the counterfactual methodology, using R to make inferences about a target who is not in the data (the counterfactual world) is considered to be inference based on generalizable statistical knowledge. On the other hand, inference about a target who *is* in the data (the actual world) depends on both generalizable knowledge and the use of their records in creating R. Thus comparison of the two attempts to isolate the specific risk due to being in the data from learning something new about the population. On the other hand, if the target person is in the data, absolute disclosure risk does not attempt to distinguish between whether an inference about sensitive information s of the target is primarily due to generalizable statistical knowledge or the target person being in the data. This harms data utility since absolute disclosure risk methodologies would limit or suppress both types of inferences.

These are not just theoretical concerns. The inability of absolute disclosure risk to factor out generalizable inference can harm demographic analysis. For example, suppose that the self-reported race and ethnicity of individuals (or of minority populations) needs to be protected. There are significant amounts of homogeneity in some block groups, tracts, counties, and even states—regions with sufficient populations for generalizable inference. For example, in 2010, 89% of the population of Montana was "White Alone," but there are clusters of minority populations in its counties, tracts, block groups, etc. Knowing that someone lives in Montana, or one of the homogeneous smaller regions, would allow race inference with at least 89% accuracy—this is the kind of statistical inference that demographers want to be able to make. Under absolute disclosure risk assessments, this accuracy would be considered a threat and the demographics of these regions would need to be suppressed. However, anyone who is aware that suppression was

done for homogeneous regions (e.g., regions with 89% homogeneity) would know these regions are homogeneous and would likely be able to guess the prevalent race/ethnicity combinations. Thus, additional complementary suppression would be needed to add uncertainty about why a region's statistics were suppressed. On the other hand, under counterfactual reasoning or careful prior-to-posterior comparisons, this would not be a risk. In particular, the counterfactual methodology seeks to protect how an individual differs from their reference population. In other words, once you know that Montana is 89% White Alone, the counterfactual approach would seek to protect inferences about how any individual differs from this baseline.

Prior-to-posterior comparisons can be set up to remove the component of inference that is due to generalizable statistical knowledge [65]. One way to do that is to have the attacker use the true data-generating distribution in their attack. However, since the true distribution is generally unknown even to the data custodian (who would need it to simulate the attack), an alternative is to use attackers who have complete confidence in their beliefs about data distributions (i.e., they do not put priors on hyperparameters [44]). In this way, the only difference between the prior and posterior distributions would be due to the use of the target person's record as part of the input to M.

4.8 Support composition

Composition is well-studied in differential privacy [36]. The counterfactual methodology strongly supports this property. On the other hand, much less is known about composition in the prior-to-posterior framework [83] and no known instances of the absolute disclosure risk framework satisfy composition. This is an area where more research is needed. Lack of composition could lead to undesirable situations in which a planned data product must be scrapped because it interacts with previously released data products, even though each of the data products were assessed to have low disclosure risk on their own. Composition has significant practical importance to statistical agencies.

4.9 Acknowledge multiple types of attackers and auxiliary information

The counterfactual framework [57] and prior-to-posterior comparisons [65] defend against multiple attackers. However, it is problematic for absolute disclosure risk. As discussed in Section 44.3, nearly every individual can have high absolute disclosure risk when the attacker is their neighbor. The disclosure may not be due to the release of R, but the absolute disclosure risk methodology does not compare any *changes* in beliefs that result from R. Thus, consideration of multiple attackers under the absolute disclosure risk framework could result in no data being released at all, or it could be compromised by the removal of many realistic attack scenarios. This again motivates using the other methodologies to make decisions about the privacy utility trade-offs when considering mechanisms for creating public data products.

4.10 Resist brittleness

For mechanisms like principled suppression [99], if a few individuals reveal their data, it can dramatically reduce the privacy protections for other individuals. This is an example of brittleness. Counterfactual comparisons assign lower disclosure risk to mechanisms that are not brittle. In fact, one of the guarantees of differential privacy is that even if all other individuals reveal their data, the target individual is still protected [36]. Not every type of prior-to-posterior and absolute disclosure risk approach resists brittleness. However they can be made to support this property [68] by adding information about other individuals into the attacker's information set A.

Brittleness also involves situations where some unperturbed statistics about the data have been released. One example is the release of the as-enumerated state population counts in the 2020 Census [7]. It is commonly believed that this irreparably breaks the differential privacy guarantees [53], but this is not the case—all it does is change the type of counterfactual analysis that is possible. To analyze the situation where the state population totals are invariant, one would use the following setup. In the counterfactual world, the target individual would submit their correct state, but would be free to alter everything else about their response. One would then compare the inference about the target in the world where their responses were truthful to the inference in this counterfactual world. Once the counterfactual world is properly chosen, the technical details of the analysis follow the examples in [64].

4.11 Be resilient to misspecification

Disclosure risk assessment methodologies rely on attacker posterior distributions $P(\cdot \mid A, M, R)$, which require specifying the attacker's prior belief in the data-generating distribution: $P(\mathcal{D})$. If one misspecifies this prior in a disclosure risk analysis, and if the true distribution differs from the specified distribution, how can one quantify the error in the disclosure risk assessment? In the counterfactual framework, the most common variations of differential privacy are immune to this problem because they provide guarantees for all attacker priors [57, 64]. In the prior-to-posterior framework, this issue is not well-studied, but Song et al. [83] showed that, in some cases, the effect of misspecifying the prior can be quantified using the Wasserstein distance between the true and specified distributions. Results for absolute disclosure risk are currently unknown.

4.12 Control computational effort

Reducing computational burden without decreasing the quality of the disclosure risk assessment should be an ongoing area of research. All the methodologies can be computationally demanding, especially for large-scale datasets. This is particularly true for methodologies that do not support composition (e.g., absolute disclosure risk and most prior-to-posterior comparisons) since they must include analysis of external datasets and previously released data products in order to avoid underestimating disclosure risk. Such analyses may already be infeasible due to the vast quantities of available data.

4.13 Control manual effort

All the methodologies require significant manual effort/technical skills, but, aside from statistical sophistication, where this effort is spent may differ. For the counterfactual framework, it is necessary to understand the complexities of differential privacy, and non-trivial effort is required in the design of mechanisms. However, there is also research into automating the design of mechanisms, known as *program synthesis*, and ensuring that the disclosure risk assessment is mistake-free, known as *verification* [82]. For variations of the prior-to-posterior comparisons that support composition, the manual effort is generally the same. For the other variations, the staff at a statistical agency would need to constantly find new sources of data that would be available to an attacker, keep up-to-date with the latest types of attacks, and develop their own attacks as well. Similar requirements apply to the absolute disclosure risk methodology.

	AbsLink	Abs	Pr2P	Cf
2.1: **Transparency**	?	?	y	y
2.2: **Universality**	n	y	y	y
2.3: **Uninformative Statistics**	NA	n	y	y
2.4: **Arbitrary Analyses** (with remote access servers)	?	?	?	?
2.5: **Arbitrarily Many Analyses**	Mathematically impossible			
2.6: **Avoiding Reconstruction**	n	y	y	y
2.7: **Generalizable Inference**	n	n	y	y
2.8: **Composition**	?	?	?	y
2.9: **Multiple Attackers**	n	n	y	y
2.10: **Resists Brittleness**	y	y	y	y
2.11: **Misspecification Resilience**	?	?	y	y
2.12: **Control Computational Effort**	n	n	?	?
2.13: **Control Manual Effort**	?	?	?	?

Table 1: Comparison of Risk Assessment Methodologies

4.14 Summary

Table 1 summarizes our analysis. We abbreviate the methodology names: absolute disclosure risk with linking is *AbsLink*, absolute disclosure risk without linking is *Abs*, prior-to-posterior comparisons are *Pr2P*, and the counterfactual methodology is *Cf*. We use y to denote that a methodology has a proof of concept with that particular desideratum, n to indicate that it is not supported at all, and ? to indicate that it may be supported but a proof of concept is needed in order to demonstrate support—hence, ? indicates potential areas of research focus. Note that it is not enough to support each desideratum with separate, incompatible proofs-of-concept; all the desiderata marked with y for a particular methodology should be supported by a common proof-of-concept. In general, we see that absolute disclosure risk methodologies have the fewest desirable properties if they are used for privacy/utility cost-benefit analysis. This is not to say that the posterior distribution that is analyzed by this methodology is not useful—this distribution is certainly informative, but only when used as part of a comparison as with the other two methodologies.

5 Discussion

Confidential responses to the census have, in the past, been misused—for example, in World War II [72]—and threats of abuse or misuse still exist. The potential for misuse is widely believed to affect data quality, possibly contributing to the undercount of Hispanic populations in 2020 [62]. Thus, a combination of access controls for the raw data and effective SDL to protect published statistics is necessary at many national statistical offices.

The efficient selection and implementation of a disclosure limitation system cannot be accomplished without a clear understanding of the intended uses for data. Independent of the degree of confidentiality protection (the privacy vs. utility trade-off), the choice of SDL approach (e.g., suppression, coarsening, or perturbation) will impact the flexibility of the resulting system to efficiently navigate that trade-off. The development of an *optimal* disclosure limitation system also requires quantifying stakeholder utility for the resulting data products, quantifying the amount of disclosure risk in the data products, and acknowledging the mathematical fact that it is impossible to provide

effective confidentiality protection while supporting every possible research purpose. (See further discussion of cost-benefit analysis in Appendix D.)

We consider three major methodologies for assessing disclosure risk: absolute disclosure risk, prior-to-posterior comparisons, and counterfactual comparisons. Of these three, absolute disclosure risk appears least useful for decision-making regarding algorithms for SDL, since it lacks important properties for these risk assessments, summarized in Section 4.14. Continuing the analogy of [53] that relates absolute disclosure risk to the chances of a patient dying, one would generally not make decisions about treatment options based on the overall probability of mortality, but rather on how the treatment would alter this probability—in other words, a comparison is needed. Thus, absolute disclosure risk is certainly an important quantity, but only useful for risk assessment when compared to something else, which is the purpose of the prior-to-posterior and counterfactual methodologies.

Hotz et al. [53] criticized the disclosure avoidance system for the 2020 Census and called for a moratorium on using differential privacy to assess the privacy risk of data products. Some of their criticism gives a path for further research in differential privacy. For example, a 10x increase in an attacker's confidence about an attribute inference may be acceptable if the initial confidence was very small (e.g., 0.0001) and less acceptable if the initial confidence were larger. However, much of their criticism rests on the adherence of differential privacy to the mathematical limitation placed on *all* disclosure limitation methodologies. This is not the fault of differential privacy, but rather a fundamental law [27]—some utility must be sacrificed to protect confidentiality, and all possible research uses cannot be simultaneously supported.

Notably, none of the criticisms to date, (except [23], which compares non-optimized versions of swapping and differential privacy) have demonstrated any technique in sufficient detail to allow its privacy/utility trade-off to be compared to that of the 2020 DAS. Thus, following scientific traditions, researchers who say "do X" or "don't do Y" need to work out a proof of concept for a sufficiently rich application, defend it, and submit it to the scientific community for analysis of weaknesses. This is especially true for something as mathematically complex as disclosure risk assessment. Additional technical challenges that need to be resolved by absolute disclosure risk methodologies can be found in SI-4.

Data users and privacy advocates can legitimately debate the policy decisions that the Census Bureau makes regarding the proper balance of privacy vs. utility and regarding the prioritization of certain use cases over others for receiving greater accuracy in the published statistics. These are public policy decisions, and the Census Bureau should rely on the expertise of its diverse external stakeholders in making them. However, the technical tools such as algorithms and disclosure risk assessment methods should be based on the best available science. A call for a moratorium on the use of certain technologies should be replaced with an exhortation to use the best researched technology that is available at the time a disclosure avoidance system must be developed and deployed.

6 Conclusions

Transparent SDL techniques help promote the trust that is necessary for surveys based on self-response. Coupled with the importance of the 2020 Census data, their use has reinvigorated interest in disclosure limitation in federal statistics because the resulting data products affect countless individuals via public policy and research. By laying out principles for the analysis of disclosure risk assessment methodologies, the aim of this paper is to help frame the debate on how the federal statistical system should proceed, and to identify the necessary research directions.

Appendices

In these supplementary materials we provide a discussion of the following items:

A. Details of how the U.S. Census Bureau engaged with stakeholders to identify major use cases for census data and assessed how differing degrees of SDL protection would impact the utility of data to support those use cases.

B. A comparison of the flexibility afforded by different SDL strategies when designing disclosure avoidance systems.

C. A discussion of synthetic data, a relevant disclosure avoidance method that did not get complete coverage in the main manuscript.

D. A discussion of cost-benefit and social welfare analysis under different disclosure risk methodologies.

E. Errors in critiques of the Census Bureau's risk analysis and differential privacy implementation for the 2020 Census.

F. Documentation of the statement (referenced in the main text of this paper) by a state demographer regarding reconstruction of microdata from published census tabulations.

G. Copy of personal communication from Gong and Meng.

A Quantifying utility via stakeholder engagement

To ensure representation of a broad spectrum of users, the agency met extensively with external stakeholder groups, including federal advisory committees; American Indian and Alaska Native tribal leaders; federal, state, and local officials; state and local demographers on the Federal State Cooperative for Population Estimates; members of the Census Bureau's State Data Center and Census Information Center networks; civil rights groups; privacy advocates; and many more. In addition to soliciting formal public comment in the Federal Register [89], the agency catalogued its efforts continuously [91].

Through these consultations, the agency developed the Detailed Summary Metrics [92], which included measures of absolute error, relative error, and outliers for a wide array of tabulations at varying levels of geography and for geographic entities of differing population sizes. For example, Wright and Irimata [100] studied the impact on redistricting and Voting Rights Act use cases. They proposed utility measures for ensuring that the largest demographic group within small places and minor civil divisions (a proxy for arbitrary, contiguous legally-defined geographic areas) with sufficient population to constitute an individual voting district, as a proportion of that geographic area's total population, was within 5% of the enumerated value at least 95% of the time.

To obtain additional feedback and identify use-cases or problems not adequately covered by the detailed summary metrics, the agency released eight sets of demonstration data products between 2019 and 2022, generated by running 2010 Census data through successive versions of the 2020 DAS [96].

This unprecedented level of public engagement for a disclosure avoidance system informed decisions about many design choices, such as the algorithm's geographic processing hierarchy, noise distributions, privacy budget allocations for different geographies and population characteristics, etc. Feedback from American Indian and Alaska Native tribal leaders about the absolute and

relative error in statistics for small tribal areas, for example, led to the incorporation of a tribal area branch of the algorithm's geographic processing hierarchy, which is the method used by the TopDown algorithm in the 2020 Census Disclosure Avoidance System to process the census data beginning with the nation as a whole, then proceeding to states, counties, census tracts, block groups, and census blocks. See [7] for details.

B A comparison of the algorithmic flexibility of SDL frameworks

Statistical agencies routinely use a risk management framework to determine when disclosure risk has been sufficiently mitigated, balancing the residual disclosure risk against the societal value of releasing the data product. Central to this assessment is the resulting utility of the data after SDL has been applied. All SDL techniques impact data utility. Whether by removing data (suppression methods), reducing precision (coarsening methods), or by otherwise introducing uncertainty (so-called perturbative methods), SDL methods reduce disclosure risk by reducing the availability and/or the precision of the data to be released. The selection of one SDL method over another or one set of implementation parameters over another has profound implications for the suitability of the resulting data for particular types of analysis. For example, income can be a highly disclosive data element because many data subjects' exact incomes make them population uniques. If an agency decided to protect income by coarsening individuals' income data into ranges for a public report, then the ranges that would support an analysis of extreme poverty can be very different from the ranges that would be used to assess gender disparities in CEO compensation. There are many other use cases of income data, and the agency would need to decide on a set of ranges that would balance the trade-off between these applications. Alternatively, the agency could permit some analyses on the unaltered confidential data and apply SDL to the outputs, for example by limiting the published precision of model parameters or infusing noise into them. Either way, the agency must assess the incremental disclosure risk of the released statistical data.

In order to support many heterogeneous use cases, it is important to choose SDL frameworks that provide the greatest flexibility and to recognize which frameworks and implementing mechanisms would limit the ability of the agency to meet its goals. When considering SDL frameworks for the 2020 Census, the Census Bureau strongly weighed the data utility implications of the leading SDL methods: suppression (item, cell, or table), coarsening, data swapping (often followed by aggregation), and noise infusion.

Item, cell and table suppression have often been used to protect tabular data [99]. Items, cells or whole tables with small counts or magnitudes, as appropriate, are initially suppressed because they are deemed too disclosive, usually called *sensitive* in this context. However, because tabulations often overlap, it would be possible to recover those suppressed data by adding together and subtracting elements from other tables. This is called a *differencing attack*, one example of a database reconstruction attack. Thus, a second round of suppression, called *complementary suppression*, is also needed to completely mask the data in the sensitive cells. The combination of primary and complementary suppression is known to cause significant loss of data and was a criticism of disclosure avoidance systems in prior censuses [69]. Abowd and Schmutte [1] pointed out that primary/complementary suppression is inherently nonignorable because the selection criteria depend on the values of data in the suppressed cells. Furthermore, Chetty et al. [21] found that suppression was significantly outperformed by techniques inspired by differential privacy for the Opportunity Atlas. The root cause of these issues is that suppression does not provide much flexibility—a cell is either suppressed or not. In contrast, noise infusion provides a continuum between perfectly precise cell counts and completely masked cell counts.

Coarsening of variables on individual records presents difficulties similar to suppression. If all records are coarsened in the same way, significant information is lost. If records are coarsened in different ways, then aggregating records becomes problematic.

Data swapping is a noise infusion technique in which pairs of individuals, or pairs of households, are matched on some variables (called the sort key) and other attributes are interchanged between the matched records. For example, the match key might be total household size and the exchanged attribute might be the geographic identifier. This swap effectively puts the data from one record into the geography of the other. Swapping introduces some uncertainty into the published data but, because it is a record-level perturbation method, it does not provide a good privacy/utility trade-off. For example, consider a set of microdata that has been swapped at a rate of 10% (so one in ten records has been altered). Given any record, attackers would have at least 90% confidence that they can correctly identify sensitive attributes of the corresponding individuals. Additional knowledge about an individual may even be used to reverse some of the swapping. However, a 10% swap rate can also have a significant impact on utility and so an agency would prefer to choose a focused approach in which swapping is mostly concentrated on individuals who have a higher risk of disclosure, typically those that are considered to be "population uniques." For highly granular data collections, however, this can be difficult as the number of population uniques typically increases with the granularity of the data publication. For example, in 2010, 57% of the person records in the decennial census were unique based on their age, sex, block, race, and ethnicity characteristics [4, p. 4]. That percentage would be expected to increase significantly when uniqueness of household composition is considered. Thus, an uncomfortably large swap rate would be needed to provide meaningful levels of privacy protections if swapping is the only disclosure avoidance method being used.

Although data swapping was approved for use in the 2010 Census, the agency believed that the swap rate was insufficient to protect confidentiality without further measures. Those additional measures included coarsening, aggregation, and synthetic microdata, which were used for the tabular products in the belief that in combination with swapping they would provide sufficient protection. However, because the 2010 data products included a large number of detailed tabulations, it is often possible to undo the aggregation, especially in census blocks with low populations. The Census Bureau withheld publication of a detailed analysis of the vulnerabilities until after the 2020 Census was completed and the full 2020 disclosure avoidance system could be tested for similar vulnerabilities; however, the methods were rigorously peer reviewed [55, 56].

The use of data swapping for future censuses would require significantly larger swap rates and less detail in the tabulations that are produced. Such a system has many fewer tuning possibilities than one based on formal privacy methods.

This left noise infusion for aggregate statistics as the Census Bureau's leading SDL candidate framework for the 2020 Census. With noise infusion, one could choose the noise distribution, the scale (amount) of noise, and the statistics to infuse with noise. For example, one isn't limited to "all-or-nothing" perturbations as in suppression and coarsening. One also is not forced to work at the record level as in data swapping (which is responsible for its problematic privacy/utility trade-off). Hence noise infusion provides a great deal of flexibility and tuning parameters for the 2020 disclosure avoidance system.

C Disclosure limitation using synthetic data

Synthetic data is an orthogonal approach to noise infusion or suppression. Any disclosure avoidance method, including synthetic data, will impact data utility for certain analyses in order to provide

some confidentiality protection. Synthetic data still carry disclosure risks, as do all data protected using any disclosure avoidance method. The key question regarding the use of synthetic data is not how it compares to unfettered access to the confidential data for anyone who wants it—that is not possible for many data sets, at least not under current laws and societal demands for data privacy—but how it compares against other disclosure avoidance frameworks. Agencies and other data custodians need to ask themselves bluntly: what are the alternatives for the confidential data at hand? In this section we describe the broad use of synthetic data to support the dissemination of data products derived from confidential data and provide a risk-utility discussion.

Several agencies now use or consider synthetic data frameworks when creating public-use files from confidential data. As examples, the Census Bureau—which already has released synthetic public-use files for the Survey of Income and Program Participation [3] and Longitudinal Business Database [66]—is researching significant changes to the confidentiality protection methods for the American Community Survey, its flagship survey, to enable the use of synthetic data [41] and validation servers. The Internal Revenue Service is making a synthetic public-use file of individual tax returns [18]. The Agency for Healthcare Research and Quality has funded a project to create a synthetic version of healthcare claims data [9]. Additionally, as described in a manual on synthetic data for national statistics organizations recently authored by the United Nations Economic Commission for Europe [86], synthetic data have been used by Statistics Canada, Statistics New Zealand, the Australian Bureau of Statistics, the United Kingdom Office of National Statistics, and in the Scottish Longitudinal Study, to name a few others. The private sector also recognizes the appeal of synthetic data approaches: there are now dozens of companies advertising their abilities to create synthetic data solutions.

The Longitudinal Business Database (LBD), which contains the annual total payroll and employee size since 1975 for every U.S. business establishment with paid employees, is an informative case study on how synthetic data can enable public access to otherwise confidential data [66]. Data from the LBD derive from Census Bureau surveys and tax files compiled by the Internal Revenue Service (IRS), making the LBD subject to the confidentiality protection provisions in both Title 13 and Title 26 of the U.S. Code. As a result, no actual values for individual establishments in the LBD can be released to the public; even the fact that an establishment filed taxes—and hence is in the data set—is protected. Thus, top coding cannot be used on monetary variables as a large fraction of exact values would be released. This also suggests that swapping would have to be done at an extremely high rate, in which case the released data would be useless for any analysis involving relationships with swapped variables. Swapping also would not protect Title 13 confidentiality because so many of the magnitude variables are population uniques and other agencies, in particular the IRS, also have copies of the data. Furthermore, many variables of interest to researchers and policy makers, for example the number of employees and total payroll, have skewed distributions with thick tails even within industry classifications. For simple noise-infusion strategies, the amount of added noise necessary to disguise these observations would have to be very large, also resulting in data of limited usefulness. Among currently available alternatives, the only way that the Census Bureau and IRS were willing to release a public-use version of the LBD was to make it fully synthetic [66].

Any researcher can use the synthetic LBD—which currently includes synthetic establishments in existence some time between 1975 and 2011—from a dedicated Census Bureau website [93]. After refining the analysis and modeling strategies based on explorations with the synthetic data, the researcher can request that the Census Bureau run the analysis on the confidential data. The Census Bureau provides results of this confidential-data analysis, processed by SDL on the outputs of the analysis to manage the disclosure risks.

Another illustrative example is the Survey of Income and Program Participation (SIPP) syn-

thetic data program [94]. In this program, the Census Bureau first combines survey responses from SIPP panels with administrative tax and benefits data into a gold standard file (GSF). The integration of these sensitive values and SIPP variables creates serious risks of disclosure and consequent harm to SIPP participants. Hence, to make a public-use version of the GSF, the Census Bureau turned to a synthetic data solution. It uses sequential regression techniques to synthesize nearly every variable—a handful of variables remain at their collected values—for every respondent record in the GSF [15]. Among existing alternatives, this was the only disclosure limitation method trusted by the providers of the administrative data to protect respondents' identities and attributes, especially against linking to other available SIPP products such as the previously released SIPP public-use microdata files.

To use the synthetic version of the GSF, users submit a request for access, after which they can perform any and all analyses supported by the schema of the GSF. The SSB program also offers validation of results on the GSF. Users provide their analysis code, proven to run on the synthetic GSF, to Census Bureau staff, who then execute the code on the GSF itself and return the results. As with the LBD, the results are processed by output SDL.

As a final example, synthetic data methods have been used to protect the confidentiality of respondents in group quarters (GQ) facilities in the American Community Survey since 2007, and in the 2010 Census [70]. GQ respondents can present difficulties for other disclosure avoidance methods that depend on matching records, such as data swapping, as the characteristics of GQ respondents can vary significantly among GQ types. For example, swapping data values of individuals in a men's prison with those in a women's shelter likely would result in data of dubious quality. Synthetic data methods can flexibly accommodate GQ data, in that the typical frameworks used for synthetic data, such as sequential modeling, can accommodate models that apply to single GQ facilities or that apply to the same type of GQ across states.

As these implementations illustrate, synthetic data systems have provided researchers with access to sensitive microdata in ways that would not have been possible with the other SDL frameworks that were in existence. We next explain why synthetic data solutions were used by considering some alternative disclosure avoidance methods traditionally employed by federal agencies and other data custodians. These alternatives and the corresponding methods for assessing their disclosure risks generally do not satisfy many of the desiderata described in the main text and in Section 2 of this supplementary information; however, they represent hypothetical alternatives that one might consider when restricted to traditional disclosure avoidance methods and averse to the use of synthetic data.

Alternative 1: No access to public-use microdata files. One alternative is to deny access to public-use microdata sets, except possibly to vetted and approved users via physical or virtual data enclaves. While limiting access can manage disclosure risks, it sacrifices some of the societal benefits of broad access to microdata. In particular, it would serve as a barrier to many user groups. For example, there are 33 Federal Statistical Research Data Centers (FSRDCs) across the United States. Many U.S. residents do not live close enough to FSRDCs to take advantage of the access they provide, although there is now a virtual access option that more than 400 researchers currently use. Because the FSRDCs primarily serve the research community (by design), access restrictions could exclude broad swathes of the public, such as local entrepreneurs, government planners, students learning the skills of data science as applied to public policy and the social sciences, and citizen scientists seeking to learn about their communities. Such groups represent legitimate stakeholders in the benefits of agency data. Accessing these data in a privacy-protected way serves that interest. Thus, while we believe providing vetted and approved users and projects access to confidential data should be a component of agencies' data dissemination policies, we expect the reduction in social welfare from Alternative 1 would be too great to make it a preferred strategy for data access,

although it has long been the practice of agencies for business data dissemination.

Alternative 2: Access to microdata treated with traditional disclosure avoidance applied at low intensity. For many agencies, past and current practice is to perturb only a small amount of data, with the hope of discouraging intruders while disturbing relationships and distributions in the data as little as possible. For example, swap only a low fraction of records, e.g., $< 1\%$; suppress only a small number of values in the data; or coarsen only slightly. Unfortunately, as noted in [53] and repeatedly in the research literature on disclosure avoidance, low intensity traditional disclosure avoidance methods may simply no longer be relied upon to protect confidentiality [1, 31, 58]. For example, consider a data set with dozens of variables, including many demographic characteristics readily available in external files. As discussed in the main text, with enough variables everyone is essentially a population unique. Since swapping 1% of records means that 99% of records are left untouched, privacy attackers may not feel discouraged from linking to external files. As another example, collapsing geography to high levels, say the public-use microdata areas (PUMA) used by the Census Bureau for the American Community Survey, may not matter when a data subject's combination of characteristics still makes them unique within that PUMA. Thus, if agencies place a high value on privacy, Alternative 2 falls short.

Incidentally, light touch disclosure avoidance methods do not necessarily have light impacts on data utility. As examples, even low rates of data swapping can deteriorate confidence interval coverage substantially [28], and top-coding of the upper tails of income can distort measures of income inequality [59].

Alternative 3: Access to microdata treated with traditional disclosure avoidance applied at high intensity. A third option is to apply traditional disclosure limitation methods with high intensity, with the goal of achieving better privacy protection. For example, agencies could suppress very high fractions of values; perform swapping across all records and variables; or add noise with huge variances. We conjecture that data users would find such data far less useful analytically than synthetic data, where records are generated from models that preserve important relationships in the data. And there is support for our conjecture from the Census Bureau's own experiments with high-intensity swapping procedures [8, Table 4].

General comments on disclosure risks and data usefulness. We close this section with additional comments on the disclosure risks and the data utility associated with synthetic data.

Regarding disclosure risks, it is important to assess disclosure risks before releasing synthetic data, as they do not guarantee zero risk [53]. Disclosure risk checks were done before the release of the synthetic LBD and the synthetic SIPP data products using best known approaches at the time. For the synthetic LBD, these checks focused on attribute disclosure risk assessments, e.g., how accurately could an intruder learn the payroll of large establishments by using the synthetic data? For the synthetic SIPP data, the checks emphasized the success of record linkage of the synthetic records to the confidential records. The insights from differential privacy tell us that such checks, while informative, do not convey mathematical guarantees of privacy against broad classes of attackers. Hence, as suggested in [53], additional research is needed on assessing disclosure risks in complex synthetic data, as well as on generating synthetic data with formal privacy guarantees.

On the other hand, synthetic data sampled from stochastic models have an appealing privacy feature not present in low-intensity traditional disclosure avoidance strategies. In practical applications, the randomness in drawing synthetic values, when implemented appropriately (so that the model does not overfit and memorize some of the training data), can strengthen privacy protections, and can even be differentially private under certain conditions [97]. Intuitively, this is because, with high probability, this process creates records where the synthetic values do not exactly match the multivariate set of confidential data values. In contrast, swapping at low rates leaves nearly all released data at the confidential values, as does coarsening applied to only a small number of

variables.

Regarding data utility, synthetic data have a well-known limitation: whatever is in the synthesis models is passed on to the analysis. Undoubtedly, some analyses will not be preserved. However, one should view the utility of synthetic data not against using the confidential data, but against other disclosure avoidance methods that provide protection akin to synthetic data, such as a 100% data swap (in the case of fully synthetic data) or massive suppression or coarsening of all variables. As noted previously, high-intensity applications of these methods can seriously degrade the utility of the data relative to synthetic data, which has the potential to preserve distributional features and relationships across variables. For the LBD and SIPP examples above, the synthetic data support very high dimensional analyses, similar to public-use microdata and much broader than the two- and three-way tabular summaries published using traditional disclosure avoidance.

It is also important to emphasize that many analyses of public-use microdata, indeed perhaps the majority for some agencies' data products, involve relatively low dimensional summaries, such as estimates of means or counts by groups. Experience shows that synthesizers can enable accurate estimation for many low-dimensional analyses [66, 3]. That said, in practical applications of synthetic data approaches, it is not "all or nothing." Researchers whose sophisticated models are not likely to be preserved, such as the examples identified in [53], can gain access to the confidential data under restricted-use agreements, such as via FSRDCs.

D Cost-benefit and social welfare analysis under different disclosure risk methodologies

Hotz et al. [53] suggested that cost-benefit analysis should be applied to managing the utility-disclosure risk decision-making without offering examples or implementation suggestions beyond their preferred disclosure risk methodology. We consider the nature of cost-benefit analysis under different approaches to disclosure risk assessment. More specifically, assuming a given data publication task and a disclosure risk methodology, how would one use that disclosure risk methodology to guide publication? In this context, the data provider needs to choose *SDL framework and implementing mechanisms* that describe how the publication task will be carried out. For example, an SDL framework and mechanism might be, say, "cell suppression under a given p-percent rule" or "swapping on a given set of characteristics and with known swap rate." In the context of formal privacy, it might be "ε-differential privacy implemented with the Laplace mechanism at a known ε".

The cost-benefit analysis described by [53] suggests that the right approach is to rank different SDL frameworks and associated implementation mechanisms in terms of their net benefits across all members of society and then choose the best one. There are two basic models for this task. First, if everyone has the same ranking over SDL frameworks and implementations, then the basic task is to assess the common ranking of these SDL regimes and choose the best one: the *universally optimal SDL regime*. Second, if there are heterogeneous preferences over SDL regimes, then it is necessary to devise a social welfare criterion that balances the competing interests of different parties, in the spirit of Abowd and Schmutte [2].

For this analysis, we assume the costs in question are the disclosure risks faced by the data subjects. The benefits are the improvements in generalizable inferences about the population from which the data are drawn. For shorthand, we refer to the latter as *data quality*, with the caveat that, in general, different users can have different assessments of the quality of a given publication depending on their background information and use case. In principle, the cost-benefit analysis might incorporate other desiderata, such as the direct monetary costs of implementing different

regimes or the effects of different regimes on response rates. We note that conducting this broader cost-benefit analysis would require the ability to trade off disclosure risk against data quality in our more limited sense, making the simplified analysis a useful starting point. If a disclosure risk methodology makes our limited cost-benefit analysis impossible, it is obviously not a candidate for the broader social choice problem.

D.1 Universally optimal SDL regimes

For the statistical agency to select a universally optimal SDL regime, it must be able to unambiguously rank all regimes such that all data users and data subjects share that same ranking. Under the absolute disclosure risk criterion, selecting a universally optimal SDL regime requires extremely restrictive assumptions. In general, it requires that all attackers have the same prior distributions over the sensitive information of each individual. Otherwise, we can not meaningfully rank the costs to individuals associated with different regimes. An alternative might be, as in some implementations of the prior-posterior methodology, to estimate the absolute disclosure risk assuming the worst-case prior. Then, because the likelihood component of the posterior distribution is fully determined by the SDL mechanism, the absolute disclosure risk would satisfy the homogeneity condition required to unambiguously rank all regimes.

D.2 General social welfare analysis

Following [2], it is possible to accommodate heterogeneity in costs associated with disclosure risk and in the benefits associated with data quality. With such heterogeneity, there is generally no universally optimal regime, but as long as a social welfare function can be specified, the problem of finding its maximum value is well-specified. The main requirements are that (1) it is possible to define social welfare in terms of the data quality measure and the disclosure risk measure, and (2) each SDL regime can be associated with a data quality measure and disclosure risk measures. The problem of maximizing the social welfare function may not always have a solution, but when it does, it is achieved by balancing the marginal cost of increasing disclosure risk with the marginal benefit of increasing data quality. Even if the social welfare function is stated in terms of absolute disclosure risk, and summarizes heterogeneous preferences over that risk, the technology need not be based on absolute disclosure risk. The relevant technological constraint is the solution to the production possibilities frontier, which can be set up to accommodate parameterization by absolute, counterfactual, or prior-posterior disclosure risk. The parameters defining the best feasible trade-off between disclosure risk and data quality also enter the social welfare function, where the disclosure risk parameters are converted to the desired metric by applying each person's prior before aggregation, just as each person's data quality valuation is applied before aggregation.

We do not mean to imply that the general social welfare analysis is feasible to implement, but it can guide policy making. For example, in setting parameters for the 2020 Census DAS, the Data Stewardship Executive Policy committee consistently allocated the minimum privacy-loss budget sufficient to meet stated data quality metrics specified in support of known, leading use cases, as described in the main text and [8]. This can be interpreted as behaving as if the production possibilities frontier were the one described by the mechanisms implemented in the 2020 DAS and the social welfare function had rectilinear iso-utility curves over the privacy-loss budget and the data quality measure.

E Details on erroneous analyses of the Census Bureau reconstruction experiment and implementation of differential privacy

E.1 Errors in published critiques

In the disclosure avoidance system for the 2010 Census, confidentiality protections for the tabular data were provided mainly by the following techniques [69, 88]:

- Data swapping for the population in households.
- Partially synthetic data for the population in group quarters.
- Age coarsening for tables at the block level and race coarsening for tables at the tract and county level.
- Aggregation of records in the universe into summary tables.

Early results from the Census Bureau's reconstruction-abetted reidentification attack were reported in the course of litigation seeking to prevent the use of differential privacy by the Census Bureau [5, 4, 6]. These results showed that in an earlier version of the reconstructed data, it is possible to obtain exact-age matches to 46.48% of the records in the 2010 Census Edited File (the confidential version of the 2010 Census data) based on the attributes block, sex, exact age, race, and ethnicity. The percentage was significantly higher using approximate matching based on age within ± 1 or age binned to 38 categories supported by the block-level tables in Summary File 1. Agreement was even higher when the analysis focused on low-population blocks [6, Table 1]–these are the geographies for which reconstruction is expected to work better. They are also the types of geographies that are widely considered to have higher disclosure risk because of the prevalence of population uniques on three widely available pseudo-identifiers: census block, sex, and age. The reconstruction portion of the Census Bureau's simulated attack only defeats aggregation and does not attempt to undo the household address swapping operations.

Meanwhile, several analyses [81, 74, 75, 40, 60, 61] asked related interesting questions.

Ruggles and Van Riper [81] ask: is 46.48% really high? To answer this question, they considered how well an attacker might do by random guessing. They concluded that a random guesser would actually have a higher success rate than 46%. However, their methodology was severely flawed. Using their methodology, the highest success rate (approximately 57%) would be achieved by an attacker who guesses that everyone is a 50-year-old female [6, pp. 15–18], even though 50-year-old females were less than 1% of the 2010 population. If 50-year old females are 1% of the population, then once one has reidentified that 1%, the job is finished. The 57% "reidentification rate" the Ruggles-Van Riper algorithm estimates is actually the probability of finding a 50-year old female in a random block, or more generally, any output from their algorithm estimates the probability of finding a random person with a given sex and age in a random block. That is not a reidentification rate.

Francis [40] asks: can a simple baseline outperform the attribute inference (identification of the pair $\{race, ethnicity\}$) that was conducted with the aid of the reconstructed Hundred-percent Detail File (HDF)? However, the methodology used for reporting early results to assess the reconstruction experiment differed from the methodology Francis used to assess his results. First, his work takes the Census Bureau's results, which are grouped by block population size, and compares them to his own results, which are grouped by racial homogeneity. The resulting group-by-group comparisons are uninformative because of this misalignment. They do not allow drawing any proper conclusions. His results are also affected by whole-person census imputations (where actual characteristics are

not known, but copied from another member of the household or a neighbor) and do not account for the homogeneity in races that is introduced by these edits and the data swapping keys.

Muralidhar [74] asks: is reconstruction really as easy as the Census Bureau experiments suggest? To answer this question, he makes a small-scale reconstruction that excluded important tabulations, used an incorrect and much simpler schema, and incorrectly asserted that the existence of multiple solutions with respect to some schema was sufficient to ignore reidentification results. With his co-author, Domingo-Ferrer, [75] they assert that the reconstruction does not imply reidentification because of suppressions that accompany correctly-implemented disclosure limitation; however, the 2010 Census data published in Summary File 1 did not impose suppression, making this argument false [43].

Kenny et al. [60] ask: is high quality inference through statistical means the same as disclosure via reconstruction? They conclude that it is. This is an incorrect conclusion because disclosure avoidance must protect how an individual's data differ from statistical predictions about them. Statistically speaking, it is the distinction between the data-generating distribution and the actual realized values of the random variables. It is also the difference between leave-one-out (LOO) error estimates of a model's accuracy compared to its in-sample training set accuracy.

In the rest of this supplementary section, we describe the flaws in these methodologies in more detail. We explain how to confirm the errors of Ruggles and Van Riper using their simulation code [80]. We provide more information about the discrepancies with the baseline provided by Francis. We show that Muralidhar's analyses gives results similar to the Census Bureau's when the mathematics are corrected. Finally, we show that Kenny et al. confound leave-one-out generalizable inference with non-leave-one-out privacy violating inferences. Kenny et al. also use flawed statistics that fundamentally overstate the error in the Census Bureau's differentially private disclosure avoidance system for the 2020 Census.

E.2 Methodological errors in Ruggles and Van Riper and their simulation

When records from the published SF1 tables are matched to the CEF in the Census Bureau's or other reconstruction experiments, the matching must be one-to-one as described in the appendices to legal declarations [4]. A record in the reconstructed data may be assigned to at most one distinct record from the confidential data. For example, suppose the reconstructed data contain the records $\{A, A, B, C\}$ and the confidential data contain the records $\{A, B, D, D\}$. There would be exactly two matches: one of the A records from the reconstructed data would be assigned to the A record of the confidential data, and the B record in reconstructed data would be matched to the B record in the confidential data.

The comparison methodology developed by Ruggles and Van Riper assesses the success rate of an attacker in a different way [81, 80]–there is no attempt to perform one-to-one matching between an attacker's file and the ground truth file (which would correspond to the confidential data). Instead, a block is chosen using population-weighted random sampling. If one block has $1,000$ people and a second block has only 1 person, the first block will be $1,000$ times more likely to be selected. After a block is selected, the attacker guesses a combination of sex and age. If this combination matches *anyone* in the block then the attacker is credited with a success. The evaluation metric of [81, 80] is the probability the attacker succeeds.

In this methodology, if a block is chosen 10 times, the attacker makes 10 guesses for the block. If the block has only one 50-year old female, but the attacker guesses 50-year old female each time, the attacker is credited with 10 matches. In other words, the matching is not one-to-one as 10 records from the attacker file are matched with just a single record in the population ground truth.

Since large blocks are selected more often and since they usually have a "50-year old female"

(it was the most common combination of age and sex in 2010 Census), the attacker should always guess this combination to maximize their success probability. This attacker would be credited with a 57% match rate in the methodology of [81, 80] even though 50-year old females were less than 1% of the entire population.

The 57% estimated by Ruggles and Van Riper is the simulation estimate of the population-weighted probability that a random block contains a 50-year old female. In other words, Ruggles and Van Riper are not measuring how well an attacker's guess matches the demographics of the population, they are only measuring how likely they are to find a random age-sex combination in a population-weighted randomly chosen block.

E.3 A Guide to the Ruggles and Van Riper simulation

This flaw can be demonstrated in a repeatable way using the provided code [80]. First, download three files from their replication archive:

- censim3.f: the Fortran code that runs the simulation.
- agesex.txt: a data file containing the cumulative distribution histogram of the age and sex of the simulated population.
- blocksize3.csv: a data file containing the cumulative distribution of block population sizes.

Next, change lines 11 and 12 in censim3.f so that they refer to the locations of the two downloaded data files. In the Fortran file, lines 62-65 implement the guess of the attacker. By adding the line `hyp=146` after line 65, one forces the attacker to always guess just one age-sex combination (corresponding to 50-year-old females). Compiling and running this code results in a file called outfile.txt that tracks the successes of an attacker. It is a spreadsheet with multiple columns. The first is the block id, the second is the population of the block, the third is how many of the attacker's guesses for the block matched *someone* from the block (the attacker guesses could match the same person multiple times, contributing to this count each time). The fourth column is the number of attacker guesses for the block that failed to match *anyone* in it. The success rate from the methodology of [81] is the sum of the third column divided by the sums of the third and fourth column. Note that each run of the code may produce slightly different results, because each run creates a random synthetic population that serves as the ground truth in the simulation.

E.4 Methodological errors in the analysis of Francis

Francis [40] proposed a simple comparison baseline for predicting an individual's race and ethnicity pair from the published tabulations: if the individual's block contains 5 or more people, guess the modal (most common) race/ethnicity in that block. First, note that this baseline would have no accuracy for people who differ from the modal race/ethnicity combination, but the Census Bureau's reconstruction experiment has high accuracy for these sub-populations as shown [4, Table 6] and [Table 3][8].

The main finding by Francis is that 11% of the United States population in 2010 appeared to live in blocks that contained at least 5 people and in which the race and ethnicity were all the same. Thus, naively, one would be expected to identify the race and ethnicity of 11% of the population with complete confidence. However, the 2010 data had nearly 6 million whole-person imputations [73] and data swapping served to make blocks appear even more homogeneous [70]. Without adjusting for these factors, Francis's analysis is overly optimistic.

In contrast, the Census Bureau's reconstruction experiment treated race and ethnicity as correctly inferred only when there was a record in an attacker's dataset and the record, augmented with information from the reconstructed data, matched a "data-defined person" (i.e., not a product of whole-person imputation) in the unswapped, confidential CEF. All truly homogeneous blocks would have perfect precision in the baseline of Francis and also in the Census Bureau's experiment. However, the Census Bureau used the narrower data-defined population to avoid allowing imputations and swapping to inflate the results.

Furthermore, the Census Bureau results were tabulated by block size, whereas the experiments of Francis were tabulated by block homogeneity (the fraction of individuals in a block that share its most frequent race × ethnicity combination), making his direct comparisons inapplicable. The Census Bureau reconstruction experiment has very high accuracy for minority populations in non-homogeneous blocks [8, Tables 2 and 3]. For the same sub-populations, the baseline from Francis would achieve a precision of exactly zero.

E.5 Methodological errors in Muralidhar's analysis

Muralihdar performed a simple example of database reconstruction using a single tract from the 2010 Census. He analyzed variation in the space of its feasible solutions. We agree that this type of analysis has merit. We do not agree that the existence of positive solution variability directly undermines interpretation of reidentification statistics. The attacker using reconstructed microdata as part of a record-linking reidentification algorithm does not have access to the sensitive data, so if solution variability were an obstacle to reidentification, then reidentification rates should be depressed by solution variability. Moreover, solution variability can be upper bounded using public data. When that upper bound is zero, as it is for about two-thirds of 2010 Census blocks [50, Slide 17], then, an attacker can explicitly focus attacks on areas where they are certain that aggregation offers no additional confidentiality protection and reidentify nonmodal persons with 95% confidence [50, Slide 19]. Here we elaborate these arguments while showing that Muralidhar used an incorrect version of the 2010 Census data schema, selected an unrepresentative block to study, and incorrectly concluded that the 2010 data schema used in our reconstruction was too coarse to permit reidentification.

Reconstructing microdata from aggregated tabular data releases is accomplished by converting the published tables into a system of integer-valued linear equations, then solving for the set (or sets) of individual-level records that conform to the constraints imposed by those equations. The more source tables that are included in the reconstruction, the greater the number of constraints that a feasible microdata reconstruction must satisfy. Muralidhar [74] attempts a single-tract example of the reconstruction of 2010 Census microdata in order to assert that the results of the Census Bureau's simulated reconstruction-abetted reidentification are less concerning than the Census Bureau's Data Stewardship Executive Policy (DSEP) committee considers them to be. In doing so, he asserts that the reconstruction process typically yields a "very large" number of possible solutions—different sets of reconstructed microdata records that all satisfy the established constraints. His example involves the reconstruction of microdata records for a single census tract (5.01) in Laramie County, Wyoming.

As described briefly in [50], the Census Bureau did compute solution variability upper bounds for all blocks in the 2010 Census using the 38-bin age coarsening supported by the block-level tables. This solution variability for the block-level-schema upper-bounds tract-level solution variability using single year of age because if the solution variability on the 38-bin age schema is zero, then the solution variability for the exact-age schema is restricted to the range of the very narrow bins (except for 85+) in the 38-bin age coarsening. All the other variables (block, sex, race, and

ethnicity) would still have zero solution variability.

The schema used in block-level data has 9,576 possible combinations of sex (2) × age bins (38) × race (63) × ethnicity (2). The sex, age, race, and ethnicity bins used in this schema are those supported by tables P1, P6, P7, P8, P9, P10, P11, P12, P12A-I, and P14 in the 2010 Census Summary File 1 (SF1) [88]. The schema used in the tract and block-level data has 2,884 possible combinations of sex (2) × age bins (103) × race (7) × ethnicity (2) at the tract level and the same 9,576 possible combinations at the block level. Therefore, the full schema supported in tract-level data has 25,956 possible combinations of sex (2) × age bins (103) × race (63) × ethnicity (2). The sex, age, race, and ethnicity bins used in the combined schema are those supported by the list above plus tract-level data in PCT12 and PCT12A-O [88]. Tract-level tables of sex × age × race × ethnicity are less granular for race and ethnicity compared with block-level tables of the same variables; whereas, block-level tables are less granular for age. Combining the tract and block-level tables into the set of constraints used in reconstruction allows results to be computed using both the block-level schema and the more detailed full tract and block-level schema. However, regardless of the schema used to report the results, exactly the same schema was imposed on target linking files when doing reidentification record linkage. Accuracy results reported in the sources cited in this section show the reidenfication based on linking using the block-level schema only was highly successful [8, Tables 2 and 3] We now relate those results to Muralidhar's critique.

Muralidhar's criteria for evaluating the reconstruction is to assume that only full precision of the age variable (to single year of age at the block level) constitutes success, even when that level of age precision is not necessary to isolate small populations and accurately link reconstructed records to an external data source with names. Muralidhar also makes two significant errors in describing the reconstruction schema: first, he asserts that "At the block level, the Age variable is grouped (except for ages 20 and 21)," but SF1 Table P14 also publishes single-year-of-age counts at the block level for ages 0, 1, 2, ..., 19, so that single-year-of-age tabulations are published at the block level up to age 21. Second, Muralidhar describes only 7 levels of race, and treats persons reporting more than one race as a single category, but SF1 Tables P8, P9, P10, and P11 publish counts for all 63 possible (non-empty) combinations of self-reported race interacted with both values of Hispanic or Latino ethnicity. These discrepancies are important, because excluding this extra granularity makes the schema seem artificially coarse, which reduces the extent to which it isolates individual records as unique or nearly unique. Using an incorrect schema also affects the number and character of distinct solutions, although the relationship here is more complicated: for a fixed total population, larger schemas tend to feature more counts that are exactly 0, which reduces the number of distinct solutions, but larger schemas also increase the number of variables in the linear relaxation of the integer program, which can cause more solutions to exist. A principled approach is required to properly account for these two effects.

Despite incorrect use of an overly simple schema, Muralidhar's focus on the variability in feasible solutions to the microdata reconstruction problem is an important aspect of bounding the extent to which aggregation alone can provide protection against privacy attacks. The existence of multiple, differing, feasible reconstructions can increase an attacker's uncertainty that they have correctly reidentified a particular data subject. But the Census Bureau's upper bound of zero solution variability for two-thirds of all blocks means that a very large component of the reconstructed data have no solution variability at all.

Muralidhar's use of an incorrect schema leads him to make other significant errors. He claims, for example, that there is no solution variability at the tract level and above. Since there are no tract-level tables that publish the full schema possible when combining tract and block-level tables, as discussed above, solution variability can remain in reconstructions based only on tract-level tables but using the full schema. In unpublished calculations of solution variability at the tract-level for a

limited number of tracts, it appears that some tracts do have positive solution variability, although the focus of the study for the full U.S. population was block-level schema solution variability.

At the block level, after the schema used by Muralihdar is corrected to include all the 9,756 possible combinations, there are still sometimes multiple distinct feasible reconstructions, but in most blocks, there is provably a single feasible reconstruction. But since two-thirds of all census blocks have no solution variability, and since the average inhabited 2010 Census block contained 50 persons, it should not be surprising that most blocks in the 2010 Census have substantially less solution variability than the very large block that Muralidhar singled-out in his illustration (block 4000 in tract 5.01 of Laramie County, Wyoming had 450 people) [87]. For 935 tracts, every single block within the tract shows zero solution variability.

Quantifying solution variability does not require access to any confidential data for verification. The attacker would know that they have the only possible set of reconstructed microdata for such blocks. The attacker would know that population unique persons in the blocks with no solution variability are population unique persons in the confidential data, an inherently disclosive attribute [29, pp. 42-43]. Because it is public information that the tables used in the reconstruction were calculated from the Census Bureau's internal swapped file (Hundred-percent Detail File), the attacker would know that they had an exact image of all records in that file, for that block, and for the schema in use. That is, the attacker would know that the publication of separate, low-dimensional tables offered no additional confidentiality protection. It is as if a single table with the full cross-tabulation of variables in the schema had been published. Only homogeneity of large populations and swapping would remain in effect as protections, and, for small populations, the attacker would know that only swapping could possibly be claimed as a source of disclosure protection.

No reidentification study is required to establish that the SF1 tables once converted to microdata violate the disclosure avoidance rules that were in place for the 2010 Census. Reidentification studies only confirm that the Census Bureau was correct in holding the official public-use microdata sample to tighter disclosure rules than swapping, and group quarters synthetic data, age-coarsening, and aggregation used for the tabular summaries. The problem is that too many exact tabular summaries were released, which enabled accurate microdata reconstruction, rendering the release of tabular summaries substantially equivalent to the release of 100% of the microdata for public-use and undermining the distinction between disclosure rules for these two kinds of products. Accurate reconstruction, in turn, enabled reidentification of very small populations, including population uniques, about which inference is clearly not generalizable, but specific to the person who is population unique.

One reason that at least two-thirds of blocks have no solution variability in the Census Bureau's reconstruction is the extreme granularity of the published 2010 Census data products and the inclusion of block-level invariants for total population and voting-age population. In particular, the block-level population invariant ensures that there is no solution variability on the block identifier for 100% of the blocks regardless of the set of tract and block-level tables used for the reconstruction. Similarly, the block-level voting-age population invariant ensures that there is no solution variability on whether the individual is an adult or minor for 100% of the blocks.

That at least two-thirds of blocks have no solution variability also refutes Muralidhar's concluding exhortation that "[i]t is up to the Census Bureau researchers to show either that the reconstruction was unique or, at the very least, that most reconstructions lead to the same conclusion regarding the identity of a respondent. Without such proof, claims of confirmed reidentification are highly suspect and easily refuted" [74]. Successful reidentification does not require that the entire reconstruction be unique, although low solution variability can be a dangerous supplement to an existing reconstruction-abetted reidentification attack by enabling an attacker to identify regions

where they can remove an important source of uncertainty. Moreover, the Census Bureau would not need to demonstrate low solution variability. Any capable attacker would be able to determine it for themselves, including identification of the blocks with provably zero solution variability in the block-level schema.

E.6 Methodological and conceptual errors in papers by Kenny et al

Kenny et al. [60, 61] evaluated the impact of demonstration data products released by the Census Bureau computed on 2010 Census data using differentially private mechanisms on redistricting outcomes. They further attempted to measure the disclosiveness of those publications.

Specifically, using the data demonstration products derived from 2010 Census data, they evaluated their ability to predict race and ethnicity from voter registration and census data. The ground truth for evaluation of prediction consisted of North Carolina voter registration records from 2021 that contained surnames, block, and race/ethnicity data.

E.6.1 Their methodology cannot measure disclosure risk

Kenny et al. cannot, by design, measure disclosure. The meaning of "disclosure" is revealing information about an individual that was provided to the agency in a confidential response. If a dataset of records from North Carolina in 2010 enables accurate inference about an individual whose information was not part of the 2010 North Carolina Summary File 1 (e.g., a 2021 North Carolina resident who lived elsewhere in 2010), then that information generally cannot be considered disclosed by the agency (exceptions exist, such as with genetic data of relatives, but such data were not part of the census).

The fact that high quality inference is possible about many North Carolina residents who lived elsewhere in 2010 means that the information they supplied in 2010 (if any) was not used in making this prediction. Hence it was not disclosed. In other words, the prediction of a person's race was based purely on data from *other* people.

For statistical models, the ability to use other people's information to make inferences can be evaluated using leave-one-out cross validation inference because, by definition, the person being predicted is not in the training data. To measure the extra risk of being in the training data, one should compare the in-sample prediction accuracy (accuracy on training data) to the leave-one-out cross-validation accuracy.

Leave-one-out statistical inferences are the method taken from robust statistics that best captures the formalism required to use the counterfactual posterior-to-posterior comparisons rather than prior-to-posterior comparisons to distinguish a generalizable scientific inference from a privacy-violating inference, as first noted in [33]. In leave-one-out inference, the counterfactual world is the one in which the target of an attack has been left out of the statistical model used for inference about or prediction of that person's data. Basically, in statistics an inference is not robust if it depends too much on the influence of one or a few observations. The simplest, but usually most computationally intensive way to do this, is to fit the model without using the first observation, then measure the error predicting that observation. This "leaves out" observation 1. Then, replace observation 1, leave out observation 2, refit the model, and measure the error predicting observation 2. Repeat until each observation has been predicted from a model where it was left out of the estimation. The cumulative error from such an algorithm estimates the leave-one-out error of the model, also called the out-of-sample prediction accuracy of the model. If the cumulative leave-one-out error of the model is much larger than the cumulative error generated when all observations are used, the inferences from the model are very dependent on one, or a few, observations. Put

differently, if the out-of-sample prediction accuracy is much worse than the in-sample prediction accuracy, then the model is not robust—it depends on one, or a few, influential observations. Proper assessment of whether an inference is scientific and generalizable or privacy violating is precisely the assessment of whether the inference is robust, in the leave-one-out sense, or not when the target person is left out of the estimation in the counterfactual world. Robust inferences or predictions are generalizable. Non-robust inferences are privacy violating.

E.6.2 Disclosure limitation should not prevent high-quality generalizable inference

Although the methodology of Kenny et al. is not measuring disclosure risk, they do raise the interesting question of whether high-quality inference should be prevented. They argue that privacy violations should be measured as the comparison between inference about an individual after a data release to the inference before the data were released. This is the prior-to-posterior methodology described in the main text. As we note in the main text, it is often undesirable to have a disclosure limitation system limit prior-to-posterior inferences, because such limitations would also prevent legitimate scientific findings (such as a dataset demonstrating the causal relationship between smoking and cancer, thus causing a change between prior and posterior distributions). Furthermore, except in rare circumstances, limiting the changes between prior and posterior distributions is known to be impossible [35]. Indeed, the technical justification that Kenny et al. rely on is a result of Gong and Meng [46] in which such restrictions were in place but not apparent due to notational issues [47]. See this supplement Section 7, which clarifies the relevant theorem.

Referring back to the analysis of Kenny et al., if one desired to make high-quality race-ethnicity inferences impossible, then block-level race/ethnicity totals would have to be made highly inaccurate, leading to low-quality redistricting data. This results in a "have your cake and eat it too" situation because one cannot make the criticism that race-ethnicity prediction is too good and yet require that redistricting results must match the confidential data.

E.6.3 High-quality generalizable inference is not a disclosure

Kenny et al. [60, 61] also claim that high quality inference through statistical means is the same as disclosure via reconstruction. This conclusion is not warranted. One reason that they are not the same is due to *calibration*–the confidence they assign to predictions, also known as reidentification precision. They report a roughly 90% accuracy in predicting race and ethnicity. However, with reconstructed data, one can focus on situations where the inference about whether the record is in the confidential data is guaranteed to be 100% accurate, relative to the underlying confidential data used in tabulation, the HDF, due to low solution variability. What is important is the difference in prediction precision between populations likely to have high precision due to generalizable inference (modal race-ethnicity groups in a census block, for example) and those likely to fail the leave-one-out generalizable inference criterion (nonmodal race-ethnicity groups in a census block, for example). In the former case (modal) a high-quality inference is often a legitimate scientific one, but in the latter case (nonmodal) the high-quality inference is clearly privacy violating because is not possible without using the target individual's actual race-ethnicity data from the census response. A consequence of this point is that high quality reconstruction allows inferences on how a person differs from statistical population characteristics (e.g., what one would predict based on what one knows about population characteristics) while high quality inference in the sense of Kenny et al. does not have this property. This is a particular concern in disclosure avoidance.

For example, consider a mixed-race household in which everyone shares the same surname (due to name changes after marriage, say) and lives in a block with only one household. Reconstruc-

tion would pose a risk if the household considered their mixed-race status to be confidential, while high-accuracy leave-one-out statistical inference would and should fail in this situation (there are no more data to use) but achieve high accuracy elsewhere (in blocks with many mixed-race households, say). In the sensitive population case (e.g., nonmodal persons in a census block), high accuracy reconstruction is a greater threat specifically because it allows better inference about how individuals differ from population characteristics.

Furthermore, reconstruction that has low solution variability means that there is high confidence associated with any *additional* information that could be attached to a given reconstructed record. If additional tables are added to the reconstruction, more information would be revealed about records and household structures, while purely statistical methods would see their accuracy degrade. In other words, inference via reconstruction carries the risk that additional information available only on a specific census response can also be accurately inferred, whereas, when extra information is added to purely statistical inferences, they suffer as the number of attributes they must predict increases; thus, they do not necessarily carry disclosure risk beyond the attributes they currently predict.

E.6.4 The major real errors in 2020 Census redistricting data

Kenny et al. claim that the 2020 DAS TopDown Algorithm (TDA) leads to undercounts of minorities. This statement is false as demonstrated by the correctly calculated error metrics in [100, 7, 92]. It is not percentage error in the size of a minority population that determine political rights, but the magnitude of the minority population as a percentage of the total population in the political entity. Wright and Irimata [100] specifically evaluated the reliability of estimates of the largest racial and ethnic populations as a percentage of total population in political entities (census places and minor civil divisions) and statistical entities (census block groups). This ratio is the measure used in redistricting scrutiny under Section 2 of the 1965 Voting Rights Act. They found that "[e]mpirical results suggest a minimum TOTAL [population] that is between 450 and 499 people in a block group provides reliable characteristics of various demographic groups in a block group based on the TDA. Similarly, a minimum TOTAL [population] that is between 200 and 249 is observed to provide reliable characteristics for places and MCDs minor civil divisions]" [100, p. 1].

Kenny et al. completely ignore the major acknowledged sources of error in the 2020 Census: differential undercounts resulting from operational and coverage errors due to difficulty contacting hard-to-reach populations, complications of the pandemic, and the political atmosphere surrounding the census, among many other causes. The Post-Enumeration Survey for the 2020 Census [63, 51], which is designed to quantify these errors, estimated a differential net undercount of -4.99% ±0.53 for Hispanic populations [63, Table 4] and -1.92% ±0.82 net undercount in the total population of Texas [51, Appendix Table 3]. These enumeration issues are much larger than any distortions due to disclosure avoidance. Simulation studies based on estimated data from the 2010 Census Coverage Measurement Program (the post-enumeration survey for the 2010 Census), making them comparable to the analyses in Kenny et al., show that the operational and coverage errors are much more important for political entities, like counties, than disclosure avoidance errors and are of the same order of magnitude as disclosure avoidance errors for census blocks [14, 95].

E.6.5 Statistical validity of percentage changes in small populations

Kenny et al. [60] also perform incorrect statistical analysis when measuring percentage changes, their statistic of choice for impugning the accuracy of the 2020 Disclosure Avoidance System (DAS). Their calculations always omit categories where there is no change when comparing the published

2010 Census results in Summary File 1 (SF1) to the results based on applying the DAS to the 2010 Census Edited File (CEF, the unswapped version of the confidential data). They always take the published 2010 Census results as the base, meaning when that result is zero, they cannot compute a percentage change. The correct way to do this, as is well known in economics, is to use an arc percentage change. The numerator is the difference between the two numbers, in this case the output of the 2020 DAS applied to the CEF and the original 2010 SF1 statistic. The denominator is the average of the two. When the two numbers are both zero, report "no change," otherwise, report 100× the ratio of the numerator to the denominator. This measure is bounded by ±200. More importantly, this measure gives credit when the DAS does not change a zero into a positive number (by design the DAS tries to preserve zeros) and symmetrically assesses when the DAS changes a one to a zero or vice versa, which is appropriate because both SF1 and the DAS output have disclosure limitation errors. In the August 25, 2022 Demonstration Data Product, there are 15.1 billion linearly independent statistics in the ensemble of released tables (the vast majority at the block levels). Of those statistics 13.6 billion (90.6%) are zero in both the DAS output (demonstration data product) and the original SF1 tables. This calculation uses only public data and can be confirmed directly by comparing the SF1 and DAS output tables for the universe of the demonstration data product. Because the percentage change measures that Kenny et al. use cannot properly account for this accuracy, they grossly overestimate the percentage error in small population statistics. To a first approximation, all their reported percentage changes for small populations should be divided by 10.

F Reconstruction of microdata from earlier censuses

At a meeting of demographers who work closely with the Census Bureau, a member of the group made the statement cited in the main text. The meeting was not public, but the statement was recorded in the transcript. We have cited it anonymously for that reason. The editor of this journal has been supplied with a copy of that meeting transcript.

G Clarification of Theorem 1.3 in Gong and Meng [46]

Re: Congenial Privacy Questions

Ruobin Gong <rg915@stat.rutgers.edu>
Fri 10/21/2022 6:02 PM

To: Daniel Kifer <dkifer@cse.psu.edu>

Cc: Meng, Xiao-li <xlmeng@g.harvard.edu>;John Maron Abowd (CENSUS/ADRM FED) <john.maron.abowd@census.gov>;Daniel Kifer (CENSUS/CES CTR) <daniel.kifer@census.gov>;Philip Leclerc (CENSUS/CED FED) <philip.leclerc@census.gov>;James Bailie <jamesbailie@g.harvard.edu>

Hi Dan (and John, Philip),

Thanks a lot for getting in touch, which is great timing! We've been feeling the attention too and in fact are working to produce an extended journal version to our conference paper (with an addition to the team, James Bailie, cc'ed). Both James and XL are in Sweden at the moment, so I'll quickly say a few things that are not necessarily on their behalf.

I agree with you that our Theorem 1.3 as presented in the conference paper needs clarification. In particular, what we're working to provide in the extended paper digs into the meaning of the prior distribution $\pi(x_i)$ in the eyes of the analyst/attacker. We think that the confusion about our theorem arises from the fact that *in general*, this prior is *not* the marginal prior from some joint prior $\pi(x_1,...,x_n)$. We now have a better notation for it, calling it $\pi_{-i}(x_i)$, to signify that it is supposed to vary according to x_i (or rather, all but x_i — hence the $-i$), and when taken together, they need not even admit a coherent joint prior distribution for all the x_i's.

The insight that you and Ashwin provided in Pufferfish is absolutely right, in that the product prior you require would essentially correspond to the specific situation that $\pi_{-i}(x_i)$ actually all are coherent marginals to a joint prior for all x_i's. But if we allow the $\pi_{-i}(x_i)$ to vary as a function of $-i$, the prior-to-posterior result still holds, just that the prior needs to reflect how much the analyst/attacker is aware of the inherent homogeneity among the data points. (For a benevolent analyst, this is precisely their "expertise" on the population under study.) And the reason we push for the prior-to-posterior semantics as an alternative to the posterior-to-posterior is that the latter is *counterfactual* — it requires the listener to imagine a database that did not materialize, whereas the former doesn't. Our experience talking to people suggests that the non-counterfactual explanation is better received. But I definitely think that we got tripped up by our own notation when we were writing Theorem 1.3, because in DP, every data point is supposed to be taken as fixed (which is not how we're used to thinking about data points) and we're conditioning on fixed things, hence this subscript disappeared from the expression.

We would love to send you our extended journal version once it's ready. (We are actually under quite some time pressure to produce it, so it won't be too long...) We also have received really helpful comments from you and the rest of the DAS team on this earlier version, which is not lost on us. We're working to incorporate them as well.

Best,
Robin

On Oct 21, 2022, at 5:24 PM, Daniel Kifer <dkifer@cse.psu.edu> wrote:

Hi Ruobin and Xiao-Li,

I hope this email finds you well, and also hope you might be able to answer a few questions about your paper on congenial privacy? It is getting a lot of attention (for example, it appears to be the background source of information for papers by Joe Hotz and Christopher Kenny) and there is one result that appears counter-intuitive, which is the discussion of section 1.4 (statistical intelligibility of differential privacy).

In this section, differential privacy is explained as a way to bound the difference between an attacker's prior and posterior distributions. Interestingly, this is a result that contradicts a paper I had with Ashwin (https://dl.acm.org/doi/10.1145/2514689) that claimed that differential privacy only bounds prior to posterior comparisons for a very limited set of priors (e.g., product distributions).

So, I was wondering if you had more details on the steps in the proof of theorem 1.3, so that I could understand the theorem better?

The way we (the DAS team) have been discussing the differential privacy guarantees is through posterior-to-posterior comparisons instead of prior to posterior comparisons. The long details are in this paper: https://arxiv.org/abs/2209.03310, but the tldr version is:
- we look at how the posterior distribution would change if a person's record was changed.
- we argue that bounding prior to posterior changes prevents learning new things from the data (and so we argue this is not what DP does)
- we argue that bounding posterior-to-posterior changes allows learning new information while still protecting privacy (and provide technical results along those lines).

The DAS team is putting together different materials for explaining what DP means, so we are trying to figure out how to resolve results that don't appear to mesh with each other.
thanks,
Dan

References

[1] J.M. Abowd and I.M. Schmutte. Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity*, pages 221–293, 2015.

[2] J.M. Abowd and I.M. Schmutte. An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1):171–202, 2019.

[3] J.M. Abowd, M. Stinson, and G. Benedetto. Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. https://ecommons.cornell.edu/handle/1813/43929, 2006. retrieved 26/04/2023.

[4] John M. Abowd. Second Declaration, Civ. A. No. 1:21-cv-01361 (ABJ), in Fair Lines America Foundation, Inc., v. United States Department of Commerce, Appendix B. https://www2.census.gov/about/policies/foia/records/disclosure-avoidance/appendix-b-summary-of-simulated-reconstruction-abetted-re-identification-attack.pdf, August 2021. retrieved 26/04/2023.

[5] John M. Abowd. Second Declaration of John Abowd, Fair Lines America Foundation, Inc. v. United States Department of Commerce and Bureau of the Census. Case No. 3:21-CV-211-RAH-ECM-KCN. https://www2.census.gov/about/policies/foia/records/disclosure-avoidance/abowd-fair-lines-v-commerce-second-declaration.pdf, 2021. retrieved 26/04/2023.

[6] John M. Abowd. Third Declaration, Document 17-1, Exhibit A in Fair Lines America Foundation, Inc., v. United States Department of Commerce. https://www2.census.gov/about/policies/foia/records/disclosure-avoidance/17-1-abowd-decl-3.pdf, November 2021. retrieved 26/04/2023.

[7] John M Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Sexton, Matthew Spence, and Pavel Zhuravlev. The 2020 Census Disclosure Avoidance System TopDown Algorithm. *Harvard Data Science Review*, Special Issue II(Special Issue 2):42–79, jun 24 2022. https://hdsr.mitpress.mit.edu/pub/7evz361i.

[8] John M. Abowd and Michael B. Hawes. Confidentiality protection in the 2020 US Census of Population and Housing. *Annual Review of Statistics and Its Application*, 10(1):119–144, 2023.

[9] Agency for Healthcare Research and Quality. Synthetic Healthcare Database for Research (SyH-DR). https://www.ahrq.gov/data/syh-dr.html, 2022. retrieved 26/04/2023.

[10] J. Trent Alexander, Michael Davern, and Betsey Stevenson. The Polls—Review: Inaccurate Age and Sex Data in the Census PUMS files: Evidence and Implications. *The Public Opinion Quarterly*, 74(3):551–569, 2010.

[11] American Cancer Society. History of the Cancer Prevention Studies. https://www.cancer.org/research/population-science/cancer-prevention-and-survivorship-research-team/acs-cancer-prevention-studies/history-cancer-prevention-study.html, undated. retrieved 27/04/2023.

[12] Baldrige v. Shapiro. 455 U.S. 345. https://supreme.justia.com/cases/federal/us/455/345/, 1982. retrieved 12/09/2023.

[13] Raef Bassily, Adam Groce, Jonathan Katz, and Adam D. Smith. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *FOCS*, pages 439–448, 2013.

[14] William Bell and Joseph Schafer. Simulation Studies to Investigate Variation in Census Counts and in Census Coverage Error Using 2010 SF-1 Data and 2010 CCM Results. https://www2.census.gov/adrm/CED/Papers/CY22/2022-01-simulation-studies.pdf, 2022. retrieved 26/04/2023.

[15] Gary Benedetto, Jordan C. Stanley, and Evan Totty. The Creation and Use of the SIPP Synthetic Beta v7.0. https://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnicalv7.pdf, 2018. retrieved 26/04/2023.

[16] Jelke G Bethlehem, Wouter J Keller, and Jeroen Pannekoek. Disclosure control of microdata. *Journal of the American Statistical Association*, 85(409):38–45, 1990.

[17] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. https://doi.org/10.48550/arXiv.1605.02065, 2016. retrieved 12/09/2023.

[18] Leonard E. Burman, Alex Engler, Surachai Khitatrakun, James R. Nunns, Sarah Armstrong, John Iselin, Graham MacDonald, and Philip Stallworth. Safely Expanding Research Access to Administrative Tax Data: Creating a Synthetic Public Use File and a Validation Server. https://www.irs.gov/pub/irs-soi/18resconburman.pdf, 2022. retrieved 26/04/2023.

[19] R. Chetty, M.O. Jackson, T. Kuchler, et al. Social capital I: Measurement and associations with economic mobility. *Nature*, 608:108—121, 2022.

[20] Raj Chetty, John Friedman, Nathaniel Hendren, Maggie R. Jones, and Sonya Porter. The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility. https://doi.org/10.3386/w25147, 2020. retrieved 12/09/2023.

[21] Raj Chetty and John N Friedman. A Practical Method to Reduce Privacy Loss When Disclosing Statistics Based on Small Samples. *Journal of Privacy and Confidentiality*, 9(2):1–23, Oct. 2019.

[22] Jennifer. H. Childs, Aleia C. Fobia, R. King, and G. Morales. Trust and Credibility in the U.S. Federal Statistical System. https://doi.org/10.13094/SMIF-2019-00001, 2019. retrieved 12/09/2023.

[23] Miranda Christ, Sarah Rahway, and Steven M. Bellovin. Differential Privacy and Swapping: Examining De-Identification's Impact on Minority Representation and Privacy Preservation in the U.S. Census. https://doi.ieeecomputersociety.org/10.1109/SP46214.2022.9833668, 2022. retrieved 12/09/2023.

[24] Aloni Cohen. Attacks on deidentification's defenses. https://arxiv.org/abs/2202.13470, 2022. retrieved 12/09/2023.

[25] Committee on Professional Ethics of the American Statistical Association. Ethical Guidelines for Statistical Practice. https://doi.org/10.5281/zenodo.7092386, 2022. retrieved 12/09/2023.

[26] Damien Desfontaines and Balázs Pejó. Sok: Differential privacies. *Proceedings on Privacy Enhancing Technologies*, 2020(2):288–313, 2020.

[27] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '03)*, pages 202–210, 2003.

[28] Jörg Drechsler and Jerome P. Reiter. Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105:1347–1357, 2010.

[29] George T. Duncan, Mark Elliot, and Juan-José Salazar-González. *Statistical Confidentiality: Principles and Practice*. Statistics for Social and Behavioral Sciences. Springer New York, 2011.

[30] C. Dwork and R. Pottenger. Toward practicing privacy. *J Am Med Inform Assoc*, 20(1):102–8, Jan 2013.

[31] C. Dwork, A. Smith, T. Steinke, and J. Ulllman. Exposed! A survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4:61–84, 2017.

[32] Cynthia Dwork. Differential Privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, pages 1–12. Springer: Berlin Heidelberg, 2006.

[33] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, page 371–380, New York, NY, USA, 2009. Association for Computing Machinery.

[34] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *TCC*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.

[35] Cynthia Dwork and Moni Naor. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1):93–107, Sep. 2010.

[36] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3–4):211–407, 2014.

[37] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *56th Annual Symposium on Foundations of Computer Science (FOCS 2015)*, 2015.

[38] Cynthia Dwork and Jonathan Ullman. The fienberg problem: How to allow human interactive data analysis in the age of differential privacy. *Journal of Privacy and Confidentiality*, 8(1):1–10, Dec. 2018.

[39] Federal Committee on Statistical Policy. Statistical Policy Working Paper 2: Report on Statistical Disclosure and Disclosure-Avoidance Techniques. Technical report, 1978.

[40] Paul Francis. A Note on the Misinterpretation of the US Census Re-identification Attack. https://arxiv.org/abs/2202.04872, 2022. retrieved 12/09/2023.

[41] M. H. Freiman, A. D. Lauger, and J. P. Reiter. Data Synthesis and Perturbation for the American Community Survey at the U.S. Census Bureau. https://www2.census.gov/adrm/CED/Papers/CY17/2017-09-FreimanLaugerReiter-ACSSynthesisPerturbation.pdf, 2017. retrieved 26/04/2023.

[42] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *KDD*, pages 265–273, 2008.

[43] Simson Garfinkel. Legacy statistical disclosure limitation techniques were not an option for the 2020 US Census of Population and Housing. *Journal of Official Statistics*, 39(3):399–410, 2023.

[44] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition, 2013.

[45] Ruobin Gong, Erica L. Groshen, and Salil Vadhan. Harnessing the Known Unknowns: Differential Privacy and the 2020 Census. *Harvard Data Science Review*, Special Issue II(Special Issue 2):1–8, jun 24 2022. https://hdsr.mitpress.mit.edu/pub/fgyf5cne.

[46] Ruobin Gong and Xiao-Li Meng. Congenial differential privacy under mandated disclosure. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, FODS '20, page 59–70, New York, NY, USA, 2020. Association for Computing Machinery.

[47] Ruobin Gong and Xiao-Li Meng. Re: Congenial privacy questions. This supplementary information, 2022.

[48] Robert Hall, Larry Wasserman, and Alessandro Rinaldo. Random differential privacy. *Journal of Privacy and Confidentiality*, 4(2):43–59, Mar. 2013.

[49] E.C. Hammond and D. Horn. The relationship between human smoking habits and death rates: A follow-up study of 187,866 men. *Journal of the American Medical Association*, 155(15):1316–1328, 1954.

[50] Hawes, Michael B. Reconstruction and Re-identification of the Demographic and Housing Characteristics File (DHC). https://www2.census.gov/about/partners/cac/sac/meetings/2022-09/presentation-reconstruction-and-re-dentification-of-dhc-file.pdf, 2022. retrieved 26/04/2023.

[51] Courtney Hill, Krista Heim, Jinhee Hong, and Nam Phan. Census Coverage Estimates for People in the United States by State and Census Operations: 2020 Post-Enumeration Survey Estimation Report. https://www2.census.gov/programs-surveys/decennial/coverage-measurement/pes/census-coverage-estimates-for-people-in-the-united-states-by-state-and-census-operations.pdf, 2022. retrieved 26/04/2023.

[52] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John Pearson, Dietrich Stephan, Stanley Nelson, and David Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLOS Genetics*, 4(8):1–9, 08 2008.

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

[54] IPUMS International. Confidentiality page. https://international.ipums.org/international/confidentiality.shtml, 2022. retrieved 12/07/2023.

[55] JASON Group. Formal Privacy Methods for the 2020 Census. https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/privacy-methods-2020-census.pdf, 2020. JSR-19-2F retrieved 26/04/2023.

[56] JASON Group. Consistency of Data Products and Formal Privacy Methods for the 2020 Census. https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-census-data-products-privacy-methods.pdf, 2022. JSR-21-02 retrieved 26/04/2023.

[57] Shiva P Kasiviswanathan and Adam Smith. On the 'Semantics' of Differential Privacy: A Bayesian Formulation. *Journal of Privacy and Confidentiality*, 6(1):1–16, 2014.

[58] S. A. Keller, S. Shipp, and A. Schroeder. Does big data change the privacy landscape? A review of the issues. *Annual Review of Statistics and Its Applications*, 3:161–180, 2016.

[59] A. Kennickell and J. Lane. Measuring the impact of data protection techniques on data utility: Evidence from the Survey of Consumer Finances. In J. Domingo-Ferrer and L. Franconi, editors, *Privacy in Statistical Databases 2006 (Lecture Notes in Computer Science)*, pages 291–303. New York: Springer-Verlag, 2006.

[60] Christopher T. Kenny, Shiro Kuriwaki, Cory McCartan, Evan T. R. Rosenman, Tyler Simko, and Kosuke Imai. The use of differential privacy for census data and its impact on redistricting: The case of the 2020 u.s. census. *Science Advances*, 7(41):eabk3283, 2021.

[61] Christopher T. Kenny, Shiro Kuriwaki, Cory McCartan, Evan T. R. Rosenman, Tyler Simko, and Kosuke Imai. Comment: The essential role of policy evaluation for the 2020 Census Disclosure Avoidance System, 2022. https://arxiv.org/abs/2210.08383.

[62] Shadie Khubba, Krista Heim, and Jinhee Hong. 2020 Post-Enumeration Survey Estimation Report: National Census Coverage Estimates for People in the United States by Demographic Characteristics. https://www2.census.gov/programs-surveys/decennial/coverage-measurement/pes/national-census-coverage-estimates-by-demographic-characteristics.pdf, 2020. retrieved 12/09/2023.

[63] Shadie Khubba, Krista Heim, and Jinhee Hong. 2020 Post-Enumeration Survey Estimation Report: National Census Coverage Estimates for People in the United States by Demographic Characteristics, PES20-G-01. https://www2.census.gov/programs-surveys/decennial/coverage-measurement/pes/national-census-coverage-estimates-by-demographic-characteristics.pdf, March 2022. retrieved 27/04/2023.

[64] Daniel Kifer, John M Abowd, Robert Ashmead, Ryan Cumings-Menon, Philip Leclerc, Ashwin Machanavajjhala, William Sexton, and Pavel Zhuravlev. Bayesian and frequentist semantics for common variations of differential privacy: Applications to the 2020 Census. https://doi.org/10.48550/arXiv.2209.03310, 2022. retrieved 12/09/2023.

[65] Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.*, 39(1), jan 2014.

[66] S. K. Kinney, J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd. Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. *International Statistical Review*, 79:363–384, 2011.

[67] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., second edition, 2002.

[68] David McClure and Jerome P. Reiter. Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Trans. Data Privacy*, 5(3):535–552, dec 2012.

[69] Laura McKenna. Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing. https://www2.census.gov/ces/wp/2018/CES-WP-18-47.pdf, 2018. retrieved 26/04/2023.

[70] Laura McKenna. Disclosure Avoidance Techniques Used for the 1960 Through 2010 Decennial Censuses of Population and Housing Public Use Microdata Samples. https://www2.census.gov/adrm/CED/Papers/CY19/2019-04-McKenna-Six%20Decennial%20Censuses.pdf, 2019. retrieved 26/04/2023.

[71] Laura McKenna. U.S. Census Bureau Reidentification Studies. https://www2.census.gov/adrm/CED/Papers/CY19/2019-04-Reidentificationstudies-20210331FinRed.pdf, 2019. retrieved 26/04/2023.

[72] JR Minkel. Confirmed: The U.S. Census Bureau Gave Up Names of Japanese-Americans in WW II. https://www.scientificamerican.com/article/confirmed-the-us-census-b/, March 2007. online only, retrieved 12/09/2023.

[73] Thomas Mule. Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Persons in the United States. https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/methodology/g-series/g01.pdf, May 2012. retrieved 26/04/2023.

[74] Krish Muralidhar. A Re-examination of the Census Bureau Reconstruction and Reidentification Attack. https://dx.doi.org/10.2139/ssrn.4103758, 2022. retrieved 12/09/2023.

[75] Krish Muralidhar and Josep Domingo-Ferrer. Database reconstruction is not so easy and is different from reidentification. *Journal of Official Statistics*, 39(3):381–398, 2023.

[76] Helen Nissenbaum. *Privacy in Context*. Stanford University Press, Redwood City, 2009.

[77] Jerome P Reiter. Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100(472):1103–1112, 2005.

[78] L. Rocher, J.M. Hendrickx, and YA de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10, 2019.

[79] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688—701, 1974.

[80] Steven Ruggles and David Van Riper. CenSim Code accompanying the article "The Role of Chance in the Census Bureau Database Reconstruction Experiment". https://osf.io/a27e6/?view_only=02225c32b4f9432cadd996d300c5ee07, June 2021. retrieved 26/04/2023.

[81] Steven Ruggles and David Van Riper. The Role of Chance in the Census Bureau Database Reconstruction Experiment. *Population Research and Policy Review*, 41(3):781–788, 2022.

[82] Calvin Smith and Aws Albarghouthi. Synthesizing differentially private programs. *Proc. ACM Program. Lang.*, 3(ICFP), jul 2019.

[83] Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. Pufferfish privacy mechanisms for correlated data. In *SIGMOD '17*, SIGMOD '17, page 1291–1306, New York, NY, USA, 2017. ACM.

[84] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[85] M. Tschantz, S. Sen, and A. Datta. SoK: Differential privacy as a causal property. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 354–371, may 2020.

[86] United Nations Economic Commission for Europe. Synthetic Data for National Statistical Organizations. https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf, 2022. retrieved 03/10/2023.

[87] U.S. Census Bureau. 2010 Census Total Population Laramie County, WY, Tract 5.01, Block 4000. https://data.census.gov/cedsci/table?q=population%20laramie%20county%20wyoming%202010&g=1000000US560210005014000_1400000US56021000501&d=DEC%20Summary%20File%201&tid=DECENNIALSF12010.P1, 2010. retrieved 26/04/2023.

[88] U.S. Census Bureau. 2010 Census Summary File 1 —Technical Documentation. https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/complete-tech-docs/summary-file/sf1.pdf, 2012. retrieved 12/09/2023.

[89] U.S. Census Bureau. Soliciting Feedback From Users on 2020 Census Data Products. https://www.federalregister.gov/documents/2018/07/19/2018-15458/soliciting-feedback-from-users-on-2020-census-data-products, July 2018. 83 FR 34111, retrieved 26/04/2023.

[90] U.S. Census Bureau. DAS 2020 Redistricting Production Code Release. https://github.com/uscensusbureau/DAS_2020_Redistricting_Production_Code, October 2021. retrieved 27/04/2023.

[91] U.S. Census Bureau. 2020 Decennial Census: Processing the Count: Disclosure Avoidance Modernization. https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html, 2022. retrieved 26/04/2023.

[92] U.S. Census Bureau. Demographic and Housing Characteristics File Demonstration Data Product August 25, 2022: Detailed Summary Metrics. https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/02-Demographic_and_Housing_Characteristics/2022-08-25_Summary_File/2022-08-25_Detailed_Summary_Metrics.xlsx, 2022. August 25, 2022 version retrieved 26/04/2023.

[93] U.S. Census Bureau. Synthetic Longitudinal Business Database (SynLBD). https://www.census.gov/programs-surveys/ces/data/public-use-data/synthetic-longitudinal-business-database.html, 2022. retrieved 26/04/2023.

[94] U.S. Census Bureau. Synthetic SIPP data. https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html, 2022. retrieved 26/04/2023.

[95] U.S. Census Bureau. Understanding Disclosure Avoidance-Related Variability in the 2020 Census Redistricting Data. https://www.census.gov/content/dam/Census/library/factsheets/2022/variability.pdf, 2022. retrieved 27/04/2023.

[96] U.S. Census Bureau. 2010 Demonstration Data Products. https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/, 2023. all vintages retrieved 26/04/2023.

[97] Yu-Xiang Wang, Stephen E. Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. https://arxiv.org/abs/1502.07645, 2015. retrieved 26/04/2023.

[98] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[99] Leon Willenborg and Ton Waal. *Statistical Disclosure Control in Practice*. Springer: New York, NY, 1996.

[100] Tommy Wright and Kyle Irimata. Empirical study of two aspects of the topdown algorithm output for redistricting: Reliability & variability. https://www.census.gov/content/dam/Census/library/working-papers/2021/adrm/SSS2021-02.pdf, 2021. August 5, 2021 update, retrieved 26/04/2023.

[101] Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. In *AIStat 2022*, volume 151, pages 4782–4817. PMLR, 28–30 Mar 2022.