Journal of Official Statistics, Vol. 39, No. 3, 2023, pp. 381–398, http://dx.doi.org/10.2478/JOS-2023-0017

Database Reconstruction Is Not So Easy and Is Different from Reidentification

Krishnamurty Muralidhar[1] *and Josep Domingo-Ferrer*[2]

In recent years, it has been claimed that releasing accurate statistical information on a database is likely to allow its complete reconstruction. Differential privacy has been suggested as the appropriate methodology to prevent these attacks. These claims have recently been taken very seriously by the U.S. Census Bureau and led them to adopt differential privacy for releasing U.S. Census data. This in turn has caused consternation among users of the Census data due to the lack of accuracy of the protected outputs. It has also brought legal action against the U.S. Department of Commerce. In this article, we trace the origins of the claim that releasing information on a database automatically makes it vulnerable to being exposed by reconstruction attacks and we show that this claim is, in fact, incorrect. We also show that reconstruction can be averted by properly using traditional statistical disclosure control (SDC) techniques. We further show that the geographic level at which exact counts are released is even more relevant to protection than the actual SDC method employed. Finally, we caution against confusing reconstruction and reidentification: using the quality of reconstruction as a metric of reidentification results in exaggerated reidentification risk figures.

Key words: Database privacy; database reconstruction; statistical disclosure control; differential privacy.

1. Introduction

Database reconstruction seems to be the nemesis of official statistics and statistical data release as they have been known so far. According to the U.S. Census Bureau's Chief Scientist:

> This (Dinur and Nissim's database reconstruction) theorem is the death knell for public-use detailed tabulations and microdata sets as they have been traditionally prepared. (Abowd 2017; Abowd et al. 2019)

Whenever a database contains personal information on a set of respondents, data protection legislation may require the organization in charge of a database, called "controller" in the European legal parlance (GDPR 2016), to take steps to protect respondent privacy. SDC, Statistical disclosure control (Dalenius 1977; Hundepool et al. 2012) is a discipline that provides methods to this end. SDC methods operate by masking, that is, altering, the data to be

[1] University of Oklahoma Price College of Business, Dept. of Marketing and Supply Chain Management, 307 West Brooks, Adams Hall Room 10 Norman, OK 73019, U.S.A. Email: krishm@ou.edu
[2] Universitat Rovira i Virgili, Department of Computer Engineering and Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia Av. Països Catalans 26, 43007 Tarragona, Catalonia. Email: josep.domingo@urv.cat
Acknowledgments: Insightful comments by Constance F. Citro to an earlier version of this article are gratefully acknowledged. Support from the European Commission (projects H2020-871042 "SoBigData++" and H2020-101006879 "MobiDataLab"), the Government of Catalonia (ICREA Academia Prize to J. Domingo-Ferrer and grant no. 2021 SGR 00115), and the Spanish Research Agency (project PID2021-123637NB-I00 "CURLING"). is gratefully acknowledged.

protected; masking can be based on data perturbation, on reduction of detail or even on generating synthetic data that preserve some of the statistics of the original data. Depending on when masking is applied, the SDC literature distinguishes among "local" protection (where data are masked by respondents themselves before being collected), "input" protection (data are masked by the controller after collecting them and all subsequent queries are answered based on the masked data) and "output" protection (queries are computed on the true original respondent data and the query outputs are masked before being released).

The process of forming a database can take place at a certain point in time or be continuous during all the lifetime of the database. In the former case, the database is said to be static, whereas in the latter case it is said to be dynamic. In static databases, the data are first collected and then they are structured to form the database. This yields a "frozen" database which is subsequently used to answer any queries. In contrast, dynamic databases contain data that are periodically changing, with new records or even attributes being added and removed over time. Organizational and corporate databases (e.g., containing data on customers, orders, and so on.) are usually dynamic. Static databases are typical in data collected for research and certain data gathered by government agencies. Obviously, not all government data are static, but in many situations they are.

Output protection is the most convenient option for dynamic databases, as it avoids having to create masked versions of the underlying changing data. Whatever the type of protection, the level of protection achieved depends on the extent to which the data have been modified (Dwork et al. 2006). In general, the greater the modification, the greater the protection but the lesser the accuracy and hence the utility of the protected data. In particular, if the original data contain outliers or unique records, greater modification may be necessary. For further discussions on SDC and the privacy-utility trade-off, (see Traub et al. 1984; Adam and Worthmann 1989; Duncan et al. 2001; Hundepool et al. 2012).

The possibility of mounting reconstruction attacks has been known for decades, and a formal theory of reconstruction attacks was developed by Dinur and Nissim twenty years ago (Dinur and Nissim 2003). U.S. Census Bureau methodologists recently stated that such attacks are no longer just a theoretical possibility, but a practical danger. Hence, they advocate using differential privacy (Dwork et al. 2006; Dwork and Roth 2014) – DP in what follows–to protect the statistical outputs of the U.S. 2020 Decennial Census as a way to thwart reconstruction of the underlying microdata. The decision to use DP motivated a lawsuit from the State of Alabama against the U.S. Department of Commerce, basically arguing utility loss (and a delay in the data release) (Alabama 2021). This lawsuit was backed by 16 other states (Associated Press 2021), but it was recently rejected by the judges (Percival 2021), on the grounds that no damage to Alabama will be provable until the DP Census data are made available.

In April 2021, the U.S. Census Bureau published a version of the 2010 Decennial Census using their new DP-based methodology, called DAS. After studying that version, several users have expressed their concerns about the utility loss caused by DAS (Kenny et al. 2021; Ruggles and Van Riper 2022; Hotz et al. 2022; Dove 2021; Schneider 2022).

Since using a DP-based methodology to prevent reconstruction is controversial in terms of utility, it remains extremely relevant to examine the real danger of reconstruction attacks and the extent to which such attacks can be warded off by DP or other methods at a reasonable utility cost.

1.2. Contribution and Plan of this Article

In this article, we reassess the risk of the original data being reconstructed by an attacker based on the protected query outputs. We first give background on reconstruction attacks (Section 2). We then examine the protection that DP can offer against reconstruction (Section 3). After that (Section 4), we discuss the critical relevance for reconstruction of the geographic level at which exact counts are reported. In Section 5 we highlight the differences between reconstruction and reidentification: using the quality of reconstruction as a measure of reidentification risk results in exaggerated reidentification risk. Conclusions and future research lines are summarized in Section 6.

2. Reconstruction Attacks

Reconstruction attacks have been known for a long time in the literature. In their pioneering work, Denning and Schlorer (1980) showed that a poorly designed database query answering system based on *output perturbation* can easily lead to disclosure of some or even all of the database records. The tool they used was the tracker attack, a carefully crafted sequence of queries aimed at isolating and disclosing specific records.

A more formal analysis of the ability to reconstruct the contents of a database using only the outputs of queries was formulated by Dinur and Nissim (2003), hereafter DN. The main step forward is their discovery that the attacker does not even need to be careful when constructing her sequence of queries. The authors assume the database is an n-bit string, that is, it contains records each of which takes value 0 or 1. They further assume all queries to be of the form "How many records in this subset are 0's?" or "How many records in this subset are 1's?". In their setting, the response to every query is computed as the true query answer plus an error E bounded in an interval ($-B, B$) for some $B > 0$. Thus, it is clear that DN assume that protection of query outputs is performed via *output perturbation* and that the error is strictly bounded.

A database reconstruction, according to DN, is a record-by-record reconstruction of the original values such that the *distance between the reconstructed values and the original values is within specific accuracy bounds*. Thus, DN's attacks are aimed at inferring the value of *each record in the original database* with a high level of accuracy. They consider two different attackers depending on their computational power:

1. **Exponential attacker.** This attacker is able to issue all possible queries. In practice, such an adversary is only realistic for *small databases*, because, say for $n \geq 100$, it would take years or decades to issue all possible queries, even with the fastest computers available. To protect against such an attacker, the output of any query is modified by adding random noise in $[-B, B]$. If the differences between the query responses obtained on the target original database and the corresponding query responses obtained on a specific candidate database are within B, then the candidate database represents a reconstruction of the original database. DN show that, in this case, the candidate database is within distance $4B$ of the target original database, where both databases are taken as binary n-vectors. Thus, unless the value B is relatively large, the candidate database is a *good* reconstruction of the original database. DN proved that, in order to prevent such a good reconstruction, B must be non-negligible compared to n, that is, B must be $O(n)$.

2. **Polynomial attacker.** This attacker issues a number of queries that is polynomial in n, which is feasible for large databases. The database protects the query outputs by adding random noise in $[-B, B]$. Using these protected query outputs, the attacker solves a linear programming problem to reconstruct the database. DN show that, with high probability, the reconstructed database is close to the original database as long as B is within $o(\sqrt{n})$ where the "little o" notation means much smaller than $\sqrt{n}$ as n grows. Hence, to achieve protection, B must be $O(\sqrt{n})$ (with "big O") for a non-negligible proportion of queries.

Thus, DN conclude that, unless the noise added to query outputs is commensurate to the size of the database ($O(n)$ for an exponential adversary and $O(\sqrt{n})$ for a polynomial adversary), the attacker is able to recreate the database. Is a noise level at least $O(\sqrt{n})$ realistic? Consider a database of size $n = 1{,}000{,}000$. What is being required is that the answers to a *non-negligible proportion* of queries differ from the corresponding true answers by about 1,000. Note that this noise does not have to be applied to all queries. Furthermore, a perturbation of about 1,000 is relatively small compared to the size of the database and to queries that may involve several hundred thousand records. Thus, the noise level required to protect against a polynomial adversary seems affordable in many situations.

Without question, DN give very relevant insights into database reconstruction using only responses to queries. The authors give a theoretical framework that explains the reconstruction risk as a function of the adversary's computational power and the noise applied to query outputs. Yet, *providing a theoretical framework for database reconstruction does not mean that every database can be reconstructed.*

For one thing, *the results by DN apply only to output perturbation, but not to local or input protection*. This is explicitly acknowledged by DN when they mention the "CD Model":

> **The CD Model.** The database algorithm above essentially creates a "private" version of the database d', and then answers queries using d'. Note that a user may retrieve the entire content of d' by querying $q_i = \{i\}$ for $1 \leq i \leq n$, after which she may answer all her other queries by herself. This result indicates that it is in some cases possible to achieve privacy in a "CD model", where users get a "private" version of the database (written on a CD), which they may manipulate (say, without being restricted to statistical queries).

Specifically, if local or input masking are implemented, the responses to all queries are based on the masked database. Hence, *for local or input perturbation, the DN framework can only reconstruct the masked database.* Now, if the local or input masking are configured to adequately protect the original database (e.g., using RR at the respondent's or microdata SDC methods described in Hundepool et al. 2012), reconstructing the masked database should not entail disclosure of sensitive information.

3. The Performance of Differential Privacy Against Reconstruction

In Dwork (2011) and Garfinkel et al. (2019), the purported solution to the reconstruction vulnerability of output-protected data is differential privacy (DP). DP was introduced by Dwork et al. (2006) as a framework for quantifying the disclosure risk associated with answering queries based on a confidential database. Assume an adversary submits a query to the database and obtains a query response R. ϵ-DP requires that, given two databases D and D' that differ in one record, and for all subsets S of the space of query responses

$$\Pr(R \in S|D \text{ is used}) \leq e^{\epsilon} \times \Pr(R \in S|D' \text{ is used}). \tag{1}$$

Essentially, DP requires that, by observing R, it must be indistinguishable within a factor e^{ϵ} whether the database D or the database D' are being used. When $\epsilon = 0$, this requirement implies that the database in use must be completely indistinguishable when observing R. In this case, the value of the record differing between D and D' stays completely confidential in spite of R being returned to the adversary. The value ϵ is usually called "privacy budget" and it should be small for the privacy condition of Expression (1) to be meaningful: Dwork (2011) recommended ϵ to be "say, 0.01, 0.1, or in some cases, ln 2 or ln 3."

A well-known property of DP is sequential composition: if k queries are individually answered with privacy levels $\epsilon_1, \epsilon_2, \ldots, \epsilon_k$, respectively, the extant privacy level after answering all k queries is $\epsilon_1 + \epsilon_2 + \ldots + \epsilon_k$.

A relaxation of DP called (ϵ, δ)-DP has also been proposed and is defined as

$$\Pr(R \in S|D \text{ is used}) \leq e^{\epsilon} \times \Pr(R \in S|D' \text{ is used}) + \delta, \tag{2}$$

where δ is the relaxation parameter. The value of δ is often interpreted to imply that ϵ-DP is satisfied with probability $1-\delta$. But a closer comparison between Expressions (1) and (2) suffices to realize that (ϵ, δ)-DP can hold without ϵ-DP being satisfied for *any* query.

The usual procedure to achieve differential privacy is to return a query answer R that consists in the query result computed on the original data plus Laplace-distributed noise. The smaller the value of ϵ and the more sensitive the query (i.e. the larger the potential change of the query result due to the change of a single record), the greater the amount of noise required.

Dwork (2011) seems to suggest that the $O(\sqrt{n})$ accuracy provided by randomized response (Warner 1965) can be outperformed by a differentially private procedure, when she writes:

> Suppose n respondents each employ randomized response independently, but using coins of known, fixed, bias. Then, given the randomized data, by the properties of the binomial distribution the analyst can approximate the true answer to the question "How many respondents have value b?" to within an expected error on the order of $O(\sqrt{n})$. As we will see, it is possible to do much better–obtaining constant expected error, independent of n.

Yet, achieving constant error independent of n clashes with the requirement of Dinur and Nissim (2003) according to which, to prevent a (polynomial) adversary from being able to reconstruct a database based on query outputs, noise at least $O(\sqrt{n})$ is needed for a non-negligible proportion of queries. As Dwork acknowledges above, achieving $O(\sqrt{n})$ noise is precisely what randomized response does.

Hence, if a differentially private procedure offers constant error independent of n, it cannot protect against reconstruction according to DN. In fact, if a DN-adversary is allowed to submit a polynomial number of queries, say $m = O(n)$, sequential composition applies, because in general queries may be on overlapping sets of individuals. Thus, the total privacy budget ϵ must be split into chunks of ϵ/m per query. Hence, since the noise applied to each query answer is inversely proportional to its privacy budget, for the Laplace mechanism the standard deviation of the noise is directly proportional to m/ϵ and therefore $O(n)$. To summarize, if ϵ-DP is correctly applied, it protects against reconstruction because it uses $O(n)$ noise, considerably more noise than randomized response. Therefore, ϵ-DP mechanisms are likely to over-protect outputs of increasing complexity, as noted in Bach (2022).

Furthermore, when comparing RR and DP it must be noted that, even though RR can satisfy the DP requirements (Dwork 2011), RR was proposed in 1965, more than a decade before the birth of the SDC discipline and four decades before DP. In fact, RR has other properties beyond DP, such as allowing an estimation of the original distribution based on the randomized distribution.

4. The Relevace of Geography and Policy Decisions for Reconstruction

U.S. law requires that the U.S. Census Bureau not "make any publication whereby the data furnished by any particular establishment or individual ... can be identified." (U.S. Census Bureau 2021a). Any individual with unique characteristics at the lowest geographic level at which the tables are released is at risk of reidentification; that is, any cell count of 1 is exposed to reidentification.

One of the key reasons for implementing DP in the Census context is the claim that the swapping approach used to protect previous decennial censuses was ineffective against reconstruction and reidentification. Garfinkel et al. (2019) provide a simple hypothetical example using primary and secondary suppression to highlight the danger of reconstruction. It has been shown (Muralidhar and Domingo-Ferrer 2021, 2022) that even this reconstruction would have been infeasible if primary and secondary suppression had been applied in the correct way (e.g., as described in Census methodology documentation (Dupre 2020) and related SDC literature (Antal et al. 2017; UNECE 2015)). It remains however true that publishing statistics at a detailed geographic level may facilitate reconstruction. We examine this issue in what follows.

4.1. The Impact of Geography

If statistics are released in small geographies, reconstruction can be performed using simple arithmetic. In fact, no matter whether swapping or DP is used as an SDC approach to protect tables, if total counts are exactly preserved at a small geographic level, then reconstruction is feasible (see Abowd and Hawes 2022, 8). The ease of reconstruction greatly depends on how small are the geographic areas for which exact counts are reported. More precisely, in the comparison between swapping and DP on the Census 2010 data conducted by the U.S. Census Bureau:

- When implementing swapping on the 2010 Census, total population and voting age counts were held invariant (exactly reported) at the block level (Abowd 2021a, 12).

- In contrast, when implementing differential privacy on the 2010 Census, only the state-level population was held invariant. Note that in 2017, block-level exact counts had been promised: "By agreement with the Department of Justice, the Census Bureau will provide exact counts at the Census block level..." (Dajani et al. 2017).

Now, there are over six million blocks versus only 50 states plus the District of Columbia. Thus, swapping was far more constrained than DP and, as a result, more disclosive. Eliminating block-level constraints (preserving total population and voting-age population counts) for swapping might put the privacy protection afforded by swapping on the same footing as DP). A fair comparison between swapping and DP would require the U.S. Census Bureau to report the results of their reconstruction attacks applied to both swapped data and DP-protected data from the 2010 Decennial Census when exact counts are preserved at the same geographic level. This would allow comparing the protection and the utility provided by both approaches; in particular, it would be interesting to see the extent to which reconstruction on DP-protected data can be performed in the same way described for swapping (Garfinkel et al. 2019, 34).

Although the U.S. Census Bureau claims to have performed a comparative analysis of DP against swapping and suppression, no specific comparative results are available. Only the following statement is provided: "to achieve the necessary level of privacy protection, both enhanced data swapping and suppression had severely deleterious effects on data quality and availability" (Abowd 2021a, 25).

Another concern is that even relaxing from exact count preservation to consistent count preservation at several geographic levels is problematic under DP. According to Garfinkel (2019, 59), noise can be added in all geographic levels of the Census 2020 as long as consistency is maintained. To ensure this consistency, the DAS methodology developed by the Census based on DP involves several postprocessing steps (Kenny et al. 2021).

4.2. The Impact of the Privacy Budget on DP

If DP is advocated as a replacement of traditional SDC methods, the privacy budget ϵ should be specified, as enjoined in Dwork et al. (2019). Taking a very small ϵ entails unaffordable utility loss, but taking ϵ very large entails very little noise addition and offers little to no protection against reconstruction, let alone reidentification (Domingo-Ferrer et al. 2021).

In fact, recent U.S. Census documents mention ϵ values as high as 19.61 in 2021 (U.S. Census 2021b) and 39.907 in 2022 (U.S. Census 2022). Let us take the 2021 ϵ value to illustrate how little privacy it achieves (the 2022 value still achieves less). We offer two different views that lead to similar conclusions:

1. First, we use the connection between DP and randomized response (Wang et al. 2016). Consider RR for a binary attribute, so that the reported randomized answer is equivalent to the true answer with probability $p \geq 0.5$ and different with probability $1-p$. Then, for any ϵ, the disclosure risk incurred by ϵ-DP is the same incurred by RR when $p = \exp(\epsilon)/(1+\exp(\epsilon))$. Specifically, $\epsilon = 19.61$ translates to binary RR with $p = 0.99999999696$, that is, to RR reporting the original value with probability practically 1, which basically amounts to no disclosure protection being offered.

2. Alternatively, we take the Dinur and Nissim perspective. For illustrative purposes, assume that the noise added is sampled from a Laplace distribution. For $\epsilon = 19.61$, the noise is bounded in the range $[-1, 1]$ with probability higher than 0.999999997, and it is bounded in the range $[-0.5, 0.5]$ with probability higher than 0.999944825. Adding this level of noise would violate the DN requirement that the noise should be $O(\sqrt{n})$ and would allow accurate reconstruction of the data.

Worse yet, even with high values of ϵ, the utility of the released DP data can be very low in some cases, as noted in Van Riper et al. (2020), Ruggles and Van Riper (2022) and Kenny et al. (2021). This is due to: (A) sequential composition, which requires splitting the privacy budget among all released outputs that are not independent of each other (for instance, among all the cells of a table, among different geographic levels, among queries related to each other, etc.); and (B) post-processing with the Census's TopDown Algorithm (TDA), which is required in order to publish data that are consistent, integral and non-negative.

The real protection and the utility loss of the high values of ϵ being proposed should be compared to those achievable using traditional SDC methods (e.g., those employed in the 2010 Decennial Census) under the same invariance constraints.

4.3. Transparency

One of the key claims when using differential privacy is that "In turn, this allows an agency like the Census Bureau to quantify the precise amount of statistical noise required to protect privacy. This precision allows the Census to calibrate and allocate precise amounts of statistical noise in a way that protects privacy while maintaining the overall statistical validity of the data" (Abowd 2021a, 22).

In fact, this is true of any methodology, including swapping. It is possible to select the swapping parameters to: (1) include (more or less) records to be swapped, (2) the attributes to be swapped, and (3) whether the swapping is performed independently for each attribute.

Releasing the ϵ parameter used in DP, as the U.S. Census Bureau does, is certainly a step in the good direction. However, this alone does not make the protection methodology transparent. The postprocessing employed remains opaque to the users. One of the key criticisms against the swapping methodology employed until the Census 2010 was that the swapping parameter (the proportion of swapped records) was not released to the public. But the U.S. Census Bureau did release an upper bound for the proportion of swapped records. Given the simplicity of swapping, this made the procedure pretty transparent. In addition, swapping also assured that certain counts were preserved even at the block level, which afforded still greater transparency.

In our opinion, transparency is not just a matter of parameter release; it also has to do with the complexity of the approach. The more complex it is, the less transparent it is to the users. In this light, the current DP-based approach can be construed as being less transparent than simple swapping.

5. Reconstruction and Reidentification are Different

Reconstruction and reidentification are two different notions:

1. Reconstruction is only the first step in the disclosure process. Note that reconstructing from the outputs of statistical queries or from released tabulations yields reconstructed data that include no identifiers.
2. Reidentification is a second and necessary step to complete a disclosure attack. In this step, the reconstructed data are *linked* to a particular individual. To this end, the attacker needs an external data source that contains identifiers plus some attributes that can be used to link with the reconstructed data. In the worst-case scenario (most favorable to the attacker), the external data source may contain the *entire* original data with identification information. For example, this worst-case scenario makes sense if the attack is conducted by the same organization that protects the data (in order to test the quality of reconstruction).

Abowd (2021a, app. B) and Garfinkel (2019) describe the reconstruction and reidentification as follows:

- The microdata records of 308,745,538 people were “reconstructed”.
- Four external commercial databases of the 2010 US population were used, which reported “Name”, “Address”, “Age”, and “Gender” of people.
- The reconstructed records were linked to the commercial databases to obtain a linked database with “Name”, “Address”, “Age”, “Gender”, “Ethnicity”, and “Race”. 45% of records could be linked.
- The linked database was compared to the U.S. Census Bureau confidential data. It is claimed that the attack got all attributes in 38% of the linked records, or equivalently for 17% of the U.S. population.
- Hence, the authors claim reidentification of 17% of the U.S. population, although Garfinkel (2019) concedes that an outside attacker would not know which reidentifications were correct.

After that, the authors go on to criticize as flawed the protection system used in the 2000 and 2010 Censuses, which relied on traditional SDC techniques. This is used as a justification to move to formal privacy, which amounts to DP. In Garfinkel (2019) it is explained that choosing the privacy budget ϵ is a public policy choice.

5.1. *Issues with Reidentification Claims*

There are several issues with the claims in Abowd (2021a, app. B):

1. It is unclear what “reconstructing” the microdata of 308,745,538 people signifies. According to Van Riper et al. (2020), it amounted to re-generating microdata records from published census block and tract tabulations, that is, from frequency tables with attributes “Census block”, “Age”, and “Gender”). This is not true reconstruction as DN describe. Note that, in general, the re-generation of microdata from a frequency table is not unique, because a frequency table contains less information than the microdata it was computed from. Hence, just re-generating one of the microdata sets that are compatible with a certain frequency table does not qualify as reconstruction of the original data: in DN’s notion of reconstruction, the accuracy bounds are essential, and no such bounds are given for the Census so called reconstruction (Muralidhar 2022).

2. Reidentification means being able to link the records in anonymized microdata with the corresponding records in an external data set containing identifiers and covering a similar population. This is not reconstruction. Since the attack was based on microdata re-generated from frequency tabulations, proper reidentification could only be conducted from those cells with count 1. In all other cases, unequivocal reidentification is impossible.
3. It has been known at least since Sweeney (2000) that matching a database containing demographic attributes such as "Municipality of residence", "Birthdate", and "Gender" against an external database containing those same attributes plus identifiers for the same population is likely to yield a high proportion of reidentifications. In fact, Ruggles and Van Riper (2022) show using a simulation that most matches reported by the U.S. Census Bureau experiment at a block level would be expected randomly and *thus fail to demonstrate a credible threat to confidentiality*. Hence, the use of DP may not be necessary. Even if the threat to confidentiality was credible, it is unclear that the Census's new DP-based TDA algorithm offers the best protection. In Francis (2022) it is shown that race and ethnicity can be inferred with more precision and less prior knowledge from TDA outputs than from the outputs of the Census previous protection algorithm.
   An alternative has also been proposed by Ruggles (2021) to investigate the impact of reconstruction on reidentification. The idea is to first match the external database to the reconstructed census data. That yields a certain matching rate r. Then take those unmatched records from the external database and compare them by block ID and PIK (the Protected Identification Key created by the Census Bureau for each original record) to the Census Edited File (the original confidential data). Let r' be the reidentification rate resulting from this comparison. If $r \approx r'$, then database reconstruction has little or no impact on reidentification; to demonstrate that reconstruction increases the reidentification risk, r should be substantially greater than r'. The U.S. Census Bureau is yet to make this comparison.

The above issues clearly show that, rather than focusing on reidentification, the Census experiment focuses on finding (non-unique) candidate reconstructions. We show next that (mis)interpreting reconstruction as reidentification may in some situations overstate and in other situations understate the real risk of reidentification.

5.2. *Misinterpreting Reconstruction as Reidentification May Overstate or Understate the Reidentification Risk*

Recall that in the U.S. Census Bureau's "reidentification" procedure described in Abowd (2021a) and Garfinkel (2019), and summarized at the beginning of Section 5, reconstructed microdata reporting,"Gender", "Age", "Race", and "Ethnicity" are linked to an external commercial database reporting "Name", "Address", "Age", and "Gener". Thus, linkage is performed using the "Age" and "Gender" attributes. As a result of linkage, a linked database is obtained that reports "Name", "Address", "Age","Gender", "Race", and "Ethnicity".

Consider three scenarios at the block level:

- **Scenario 1.** Block whose reconstructed data consist of ten individuals with "Age" = 44, "Gender" = Male, "Race" = White and "Ethniciy" = Not_Hispanic. The commercial database contains "Name", "Address", "Age" = 44 and "Gender" = Male for all individuals in this block.
- **Scenario 2.** Same data as in Scenario 1, but with an additional attibute *Relationship*, which according to Garfinkel (2019) is also collected for each person in a block and can take 17 different values. Assume that in, this scenario, each of the ten persons in the block has a different "Relationship" value. Like in Scenario 1, the commercial database contains "Name", "Address", "Age" = 44 and "Gender" = Male for all individuals in this block.
- **Scenario 3.** Block whose reconstructed data consist of ten individuals with "Age" = 44, "Gender" = "Male", such that such that all ten of these individuals belong to different ("Race", "Ethnicity") combinations. The commercial database contains "Name", "Address", "Race" and "Ethnicity" for all individuals in this block, but no "Age" or "Gender".

In Scenario 1, the U.S. Census Bureau's procedure would yield a 100% reconstruction, because the attacker would always be able to associate the correct "Race" and "Ethnicity" to the ten names and addresses in the block. Yet, claiming that this 100% reconstruction amounts to 100% reidentification is patently incorrect, because the attacker has no way to confirm the identification of ("Name", "Address") for the ten individuals who are indistinguishable from one another – "reidentification" in this case can be attributed to the homogeneity of a block and is not a true reidentification. The U.S. Census Bureau document McKenna and Haubach (2019) states that "it is necessary to verify the proposed matches by comparing the suppressed identities in the microdata with the identities in the external data set to see if the matches are true matches or false matches."

The above point that correctly reconstructing "Ethnicity" and "Race" does not amount to reidentification becomes apparent in Scenario 2. When the attribute "Relationship" is added with different values for all ten individuals, it becomes clear that the reidentification probability for any specific individual is in fact 1/10.

In Scenario 3, both the probability of correct reconstruction and the probability of correct reidentification are 1, but for different reasons:

- Since all individuals have the same combination of ("Age", "Gender"), reconstructing the values of these attributes for the ten individuals is trivial, which yields 100% reconstruction. Note that if not all individuals had the same combination, then the probability of correct reconstruction would be less than 1.
- Since all individuals have different combinations of ("Race", "Ethnicity"), unequivocally linking each of the ten records in the reconstructed data to its corresponding record in the commercial database is straightforward, which yields 100% reidentification.

The above shows that reconstruction and reidentification are different notions. The bottom line is as follows: whereas *reconstruction is helped by homogeneity of the missing confidential attributes, reidentification is helped by heterogeneity of the quasi-identifiers through which linkage is performed.*

Hence, the reconstruction procedure described in Abowd (2021a) and Garfinkel (2019) does not yield an appropriate measure of reidentification risk. In fact, it is likely to overstate the reidentification risk (as in Scenario 1), since at the block level ("Race", "Ethnicity") can be expected to be fairly homogeneous, which makes Scenario 1 more likely than Scenario 3 (Ruggles and Van Riper 2022).

Interestingly, researchers at the U.S. Census Bureau have performed in the past extensive research in reidentification risk (e.g., Winkler 1999). To assess the true risk of reidentification, it is necessary to assume the following. At the block level, the reconstructed data consist of ("Block ID", "Gender", "Age", "Race", "Ethnicity", and "Relationship".) and the attacker has the attributes ("Name", "Address", "Block ID", "Gender", "Age", "Race", "Ethnicity", and "Relationship"). The objective of reidentification is to uniquely link a record from the (unidentified) reconstructed data to a record in the (identified) attacker's data thereby attaching ("Name", "Address") to the reconstructed data. *Such a procedure will correctly assess the reidentification risk in the scenarios described above*; as mentioned above and in McKenna and Haubach (2019), once the linkage is established, reidentification needs to be validated by checking that the linkage is unique and that identities (name and address in this case) match between the attacker's record and the original record to which the unidentified reconstructed record corresponds. Reconstructing unidentified records, in itself, does not pose a real disclosure threat. *Reconstruction* in the DN sense also requires to be supplemented by correct re-identification (Bach 2022). Only then does it constitute real disclosure.

6. Conclusions and Future Work

In this article, we have reassessed the feasibility of reconstructing a data set based on the outputs of statistical queries computed on it. The danger of reconstruction has been cited as an argument to justify the use of differential privacy in official statistics, most notably in the case of the 2020 Census of the U.S.A. Using DP, however, will most likely result in a decrease of the utility of the statistical outputs of that Census. This article has investigated to what extent reconstruction is a real danger.

We first examined the state of the art in reconstruction theory – Dinur and Nissim's framework – and we concluded that local or input protection appear as good ways to resist reconstruction. If the U.S. Census Bureau were to stick to the so-called CD-model and produced locally protected or input-protected data (e.g., using RR or microdata masking discussed in Hundepool et al. (2012), or the methods used in the 2010 Census), then reconstruction would not be a real danger: at most the attacker would be able to reconstruct the locally protected or the input-protected data, rather than the original data. Differential privacy is also an option, but it may add more noise than strictly required to counter reconstruction, thereby leading to unnecessary utility loss (Dove 2021; Hotz et al. 2022; Bach 2022), or it may offer less protection than previous approaches (Francis 2022).

We then highlighted the relevance for protection of the geographic levels at which exact population counts are preserved. No matter the SDC methods used, preserving counts in small geographies facilitates reconstruction, while not preserving counts in small geographies goes a long way towards avoiding reconstruction, but also reducing utility.

Finally, we have warned against using the amount of reconstruction as a measure of reidentification risk, which results in exaggerated reidentification risk. Whereas reconstruction requires only query outputs or tabular unidentified outputs and is favored by the homogeneity of the missing attribute values, reidentification also needs external identified sources and is favored by the heterogeneity of the values of the attributes used for linking with those sources.

An additional concern are the successive increases of the value of ϵ during the process. Increasing ϵ implies a loss of privacy. The U.S. Census Bureau started with Laplace noise addition with $\epsilon = 4.5$ and subsequently increased to $\epsilon = 10.2$. To further improve the utility of the released data, the Bureau adopted zero-concentrated DP (with noise from a discrete Gaussian distribution, (Bun and Steinke 2016)) in place of Laplace noise. The parameter ρ of zero-concentrated DP can be used to compute equivalent values ϵ and δ for (ϵ, δ)-DP. For a given ρ there are many equivalent combinations (ϵ, δ). However, for fixed δ (in the case of the Census it is $\delta = 10^{-10}$), then each ρ has a single equivalent ϵ. In 2021, this equivalent global ϵ was 19.61 (U.S. Census 2021). This value was further revised to a global $\epsilon = 39.907$ in year 2022 (U.S. Census 2022). This has a great impact on privacy. The privacy level associated with $\epsilon = 39.907$ is worse than the privacy level given by $\epsilon = 4.5$ by a factor $e^{39.907}/e^{4.5} = 2.382 \times 10^{15}$. Referring to Apple's use of $\epsilon = 14$, Frank McSherry, one of the inventors of DP, commented that it was "pointless" in terms of privacy (Greenberg 2017).

Actually, $\epsilon = 39.907$ is over 1.78×10^{11} times worse than $\epsilon = 14$. In Abowd (2021b), the then Census's Chief Scientist said about ϵ that "specifically it limits the statistical power of all possible tests for whether a particular individual's data record (or portions thereof) was used to produce a collection of statistics versus the record of another, arbitrary individual." With $\epsilon = 39.907$ used in 2022, this implies that the probability that a particular individual's data record was used can be over 2.14×10^{17} times higher versus the record of another, arbitrary individual. Since the current U.S. population is only 3.31×10^{8}, with an $\epsilon = 39.907$ any target U.S. inhabitant might be reidentifiable. In fact, this also held for the $\epsilon = 19.61$ used in 2021.

Even with this relaxation of the value of ϵ, there are still very serious utility concerns. Consider the following report in the New York Times (Wines 2022): "According to the 2020 census, 14 people live there (in Census Block 1002 in downtown Chicago) –13 adults and one child. Also according to the 2020 census, they live underwater. Because the block consists entirely of a 700-foot bend in the Chicago River." Or this analysis from Cornell University (Cornell 2021) of the 2021 Census DAS release for New York state which shows that in 6.1% of the blocks, the household population is greater than 0, but the number of occupied houses is 0; in 2.5% of the blocks, the household population is less than the number of occupied houses (which means there is less than 1 person per household); and in 0.8% of the blocks, the household population is 0, but the number of occupied houses is greater than 0. These results are impossible and would not have occurred in the 2010 Census. Thus, even with large ϵ, the differentially private noise addition procedure is not capable of providing accurate and consistent results. In fact, the U.S. Census Bureau recently announced that "for the time being, the ACS PUMS (American Community Survey Public Use Microdata Sample) data product will still be protected using traditional disclosure avoidance methods", since it is "not clear that differential privacy would ultimately be the best option." (Daily 2022)

In this study, we conclude that the concern of database reconstruction resulting in mass disclosure is unwarranted. We believe that these claims are based on a comparison that is incomplete and opaque – only the Census Bureau can assess or verify the true reidentification results. Other researchers, some of them mentioned above, have raised serious concerns regarding the accuracy and consistency of the output. Hence, it is not clear that differential privacy is the best option for the 2020 decennial census data. We suggest a comprehensive, independent, fully documented, peer-reviewed assessment of the efficacy of alternative methods.

7. References

Abowd, J.M. 2017. "Research data centers, reproducible science, and confidentiality protection: the role of the 21st century statistical agency." Presentation to Summer DemSem, Wisconsin Federal Statistical RDC. June 5, Madison, Wisconsin. Available at: https://www2.census.gov/cac/sac/meetings/2017-09/role-statistical-agency.pdf (accessed February 2023).

Abowd, J.M., I.M. Schmutte, W.N. Sexton, and L. Vilhuber. 2019. "Why the economics profession cannot cede the discussion of privacy protection to computer scientists." Presentation to The Future of Economic Research under Rising Risks and Costs of Information Disclosure. Allied Social Science Associations Annual Meetings. January 5, Atlanta, Georgia. https://ecommons.cornell.edu/handle/1813/60836 (accessed February 2023).

Abowd., J.M. 2021a. *Declaration of John M. Abowd. Case no. 3:21-CV-211-RAH-ECM-KCN.* U.S. District Court for the Middle District of Alabama. Availabe at: https://censusproject.files.wordpress.com/2021/04/2021.04.13-abowd-declaration-alabama-v.-commerce-ii-final-signed.pdf (accessed June 2023).

Abowd, J.M. 2021b. *Supplemental Declaration of John M. Abowd. Case no. 3:21-CV-211-RAH-ECM-KCN.* U.S. District Court for the Middle District of Alabama. Available at: https://www.brennancenter.org/sites/default/files/2021-06/M.D.%20Ala.%2021-cv-00211%20dckt%20000116_001%20filed%202021-04-26%20Abowd%20declaration.pdf (accessed June 2023).

Abowd J.M., and M.B. Hawes. 2022. *Confidentiality Protection in the 2020 US Census Population and Housing*. Available at: https://arxiv.org/pdf/2206.03524.pdf (accessed February 2023).

Adam, N.R., and J.C. Worthmann. 1989. "Security-control methods for statistical databases: a comparative study." *ACM Computing Surveys* 21(4): 515–556. DOI: https://doi.org/10.1145/76894.76895.

Alabama. 2021. *Alabama* v. *U.S. Dep't of Commerce. 2021*. Brennan Center for Justice. https://www.brennancenter.org/our-work/court-cases/alabama-v-us-dept-commerce (accessed February 2023).

Antal, L., T. Enderle, and S. Giessing. 2017. *Statistical disclosure control methods for harmonised protection of census data*. Deliverable D3.1 Part I, Eurostat contract

"Harmonised Protection of Census Data in the ESS", Available at: https://ec.europa.eu/eurostat/cros/system/files/methods_for_protectingcensus_data.pdf (accessed February 2023).

Associated Press. 2021. "16 states back Alabama's challenge to Census privacy tool." *U.S. News*. Available at: https://www.usnews.com/news/us/articles/2021-04-13/16-states-back-alabamas-challenge-to-census-privacy-tool (accessed February 2023).

Bach. 2022. "Differential privacy and noisy confidentiality concepts for European population statistics." *Journal of Survey Statistics and Methodology*, 10: 642–687. DOI: https://doi.org/10.1093/jssam/smab044.

Bun, M., and T. Steinke. 2016. "Concentrated differential privacy: simplifications, extensions, and lower bounds." In. Theory of Cryptography Conference-TCC, October 31–November 3, Beijing, China. Springer: 635-658. DOI:https://doi.org/10.1007/978-3-662-53641-4_24. https://link.springer.com/chapter/10.1007/978-3-662-53641-4_24.

Cornell. 2021. Census 2020 results: Data and Analyses for New York from the data products as they are released over time by the U.S. Census Bureau. Cornell Program on Applied Demographics. Available at: https://pad.human.cornell.edu/census2020/index.cfm#das (accessed February 2023).

Dalenius. T. 1977. "Towards a Methodology for Statistical Disclosure Control." *Statistisk Tidskrift* 15: 429–444. Available at: https://ecommons.cornell.edu/bitstream/handle/1813/111303/dalenius-1977.pdf?sequence = 3&isAllowed = y (accessed June 2023).

Daily, D. 2022. *Disclosure avoidance protections for the American Community Survey*. U.S. Census Bureau. Available at: https://www.census.gov/newsroom/blogs/random-samplings/2022/12/disclosure-avoidance-protections-acs.html (accessed February 2023).

Dajani, A.N., A.D. Lauger, P.E. Singer, D. Kifer, J.P. Reiter, A. Machanavajjhala, S.L. Garfinkel, S.A. Dahl, M. Graham, V. Karwa, H. Kim, P. Leclerc, I.M. Schmutte, W.N. Sexton, L. Vilhuber, and J.M. Abowd. 2017. "The modernization of statistical disclosure limitation at the U.S. Census Bureau. In Census Scientific Advisory Committee Meeting, Sepember 14–15, Suitland MD, USA. Available at: https//www.census.gov/library/video/2017/2017-09-sac.html (accessed February 2023).

Denning, D.E., and J. Schlorer. 1980. "A fast procedure for finding a tracker in a statistical database." *ACM Transactions on Database Systems* 5(1): 88–102. DOI: https://doi.org/10.1145/320128.320138.

Dinur, I., and K. Nissim. 2003. "Revealing information while preserving privacy." In. Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Jun. 9–12. San Diego CA, USA.: 202–210, DOI: 10.1145/773153.773173; https://dl.acm.org/doi/10.1145/773153.773173.

Domingo-Ferrer, J.D. Sánchez, and A. Blanco-Justicia. 2021. "The limits of differential privacy (and its misuse in data release and machine learning)." Communications of the ACM 64(7): 33–35. DOI: https://doi.org/10.1145/3433638.

Dove, I. 2021. *Applying differential privacy protection to ONS mortality data, pilot study*. Office for National Statistics. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/applyingdifferentialprivacyprotectiontoonsmortalitydatapilotstudy (accessed June 2023).

Duncan, G.T., S.A. Keller-McNulty, and S. Lynne. 2001. *Disclosure Risk vs. Data Utility: The RU Confidentiality Map.* National Institute of Statistical Sciences. Technical Report no. 121. Available at: https://www.niss.org/sites/default/files/technicalreports/tr121.pdf (accessed February 2023).

Dupre, S. 2020. *Disclosure avoidance and the Census. Select Topics in International Censuses*. U.S. Census Bureau. Available at: https://www.census.gov/content/dam/-Census/library/working-articles/2020/demo/disclosure_avoidance_and_the_census_-brief.pdf (accessed February 2023).

Dwork, C. 2011. "A firm foundation for private data analysis." *Communications of the ACM* 54(1): 86–95. DOI: https://doi.org/10.1145/1866739.1866758.

Dwork, C.N. Kohli, and D. Mulligan. 2019. "Differential privacy in practice: expose your epsilons!" *Journal of Privacy and Confidentiality* 9(2): 9–22. DOI: https://doi.org/10.29012/ jpc.689.

Dwork, C., F. McSherry, K. Nissim, and A. Smith. 2006. "Calibrating noise to sensitivity in private data analysis." In Theory of Cryptography Conference – TCC 2006, March. 4–7, New York NY USA. Springer: 265–284. DOI: https://doi.org/10.1007/11681878 https://link.springer.com/chapter/10.1007/11681878_14.

Dwork, C. and A. Roth. 2014. *The Algorithmic Foundations of Differential Privacy.* Now Publishers. DOI: https://doi.org/10.1561/0400000042.

Francis, p. 2022. "A note on the misinterpretation of the US Census re-identification attack." In *Privacy in Statistical Database – PSD 2022*, Sepember 21–23, Paris, France. Springer: 299–311. DOI: https://doi.org/101007/9783-031-13945-1 https://link.springer.com/chapter/10.1007/978-3-031-13945-1_21.

Garfinkel, S. 2019. "Deploying differential privacy for the 2020 Census of Population and Housing." In Privacy Enhancing Technologies Symposium–PETS 2019, July 16–20, Stockholm, Sweden. Available at: https://simson.net/page/Main_Page (accessed February 2023).

Garfinkel, S., J.M. Abowd, and C. Martindale. 2019. "Understanding database reconstruction attacks on public data." *Communications of the ACM* 62(3): 46–53. DOI: https://doi.org/10.1145/3291276.3295691.

GDPR. 2016. *General Protection Regulation.* Regulation (EU) 2016/679. Available at: https://gdpr-info.eu (accessed February 2023).

Greenberg, A. 2017. "How one of Apple's key privacy safeguards falls short." *Wired.* Available at: https://www.wired.com/story/apple-differential-privacy-shortcomings/ (accessed February 2023).

Hotz, V.J., C.R. Bollinger, T. Komarova, C.F. Manski, R.A. Moffitt, D.Nekipelov, A. Sojourner, and B.D. Spencer. 2022. "Balancing data privacy and usability in the federal statistical system." *PNAS* 119(31): e2104906119. DOI: https://doi.org/10.1073/pnas.21049 06119.

Hundepool, A.J. Domingo-Ferrer, L. Franconi, K. Spicer, P.-P. De Wolf, S. Giessing, and E. Schulte Nordholt. 2012. *Statistical Disclosure Control.* Wiley. DOI: https://doi.org/10.1002/9781118348239.

Kenny, C.T.S. Kuriwaki, C. McCartan, E.T.R. Rosenman, T. Simko, and K. Imai. 2021. "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. DOI: https://doi.org/10.1126/sciadv.abk3283.

McKenna, L., and M. Haubach. 2019. *Legacy techniques and current research in disclosure avoidance at the U.S. Census Bureau.* Research and Methodology Directorate, U.S. Census Bureau. Available at: https://www.census.gov/library/working-articles/2019/adrm/CED-WP-2019-005.html (accessed February 2023).

Muralidhar, K. 2022. "A re-examination of the Census Bureau reconstruction and reidentification attack. In Privacy in Statistical Database – PSD 2022, September 21–23, Paris, France. Springer: DOI: https://doi.org/10.1007/978-3-031-13945-1 https://link.springer.com/chapter/10.1007/978-3-031-13945-1_22.

Muralidhar, K., and J. Domingo-Ferrer. 2021. "Database reconstruction is very difficult in practice!" In 2021 Joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality, december 1–3, Poznan, Poland. Available at: https://unece.org/sites/default/files/2021-12/SDC2021_Day1_Muralidhar_AD.pdf. (accessed February 2023).

Muralidhar, K., and J. Domingo-Ferrer. 2022. "Census reconsiderations". *Communications of the ACM* 65(6): 11. DOI: https://doi.org/10.1145/3532630.

Muralidhar, K., and R. Sarathy. 2009. "Privacy violations in accountability data released to the public by state educational agencies." In Federal Committee on Statistical Methodology Research Conference, November 2–4, Washington D.C. USA. Available at: https://www.researchgate.net/profile/Rathindra-Sarathy/publication/273448878_Privacy_Violations_in_Accountability_Data_Released_to_the_Public_by_State_Educational_Agencies_Rathindra_Sarathy/links/5501d43e0cf231de076ca7b3/Privacy-Violations-in-Accountability-Data-Released-to-the-Public-by-State-Educational-Agencies-Rathindra-Sarathy.pdf (accessed June 2023).

Percival, K. 2021. *Court rejects Alabama challenge to Census plans for redistricting and privacy.* Brennan Center for Justice. Available at: https://www.brennancenter.org/our-work/analysis-opinion/court-rejects-alabama-challenge-census-plans-redistricting-and-privacy (accessed February 2023).

Ruggles. S. 2021. *Personal communication, November 9.*

Ruggles, S., and D. van Riper 2022. "The role of chance in the Census Bureau database reconstruction experiment." *Population Research and Policy Review* 41: 781–788. DOI: https://doi.org/10.1007/s11113-021-09674-3.

Schneider, M. 2022. "Researchers ask Census to stop controversial privacy method." AP News., Available at: https://apnews.com/article/census-2020-us-bureau-government-and-politics-20e683c71eeb62ee4b7792d7d8530419 (accessed February 2023).

Sweeney, L. 2000. *Simple demographics often identify people uniquely.* Carnegie Mellon University, Data Privacy Working Paper 3. Available at: https://ggs685.pbworks.com/wZfile/fetch/94376315/Latanya.pdf (accessed February 2023).

Traub, J.F., Y. Yemini, and H. Wozniaknowski. 1984. "The statistical security of a statistical database." *ACM Transactions on Database Systems*, 9(4): 672–679. DOI: https://doi.org/10.1145/1994.383392.

UNECE-CES. 2015. (United Nations Economic Commission for Europe – Conference of European Statisticians) *Recommendations for the 2020 Censuses of Population and Housing*. United Nations Publications, New York, NY. Available at: https://unece.org/DAM/stats/publications/2015/ECECES41_EN.pdf (accessed February 2023).

U.S. Census Bureau. 2021a. *Legal Authority and Policies for Data Linkage at Census*. Available at: https://www.census.gov/about/adrm/linkage/about/authority.html (accessed June 2023).

U.S. Census Bureau. 2021b. *Census Bureau sets key parameters to protect privacy in 2020 Census results*. Release Number CB21-CN.42. Available at: https://www.census.gov/-newsroom/press-releases/2021/2020-census-key-parameters.html (accessed February 2023).

U.S. Census Bureau. 2022. *Privacy loss Budget Allocation*. Available at: https://www2.census.gov/programs-surveys/decennial/2020/program-management/-data-product-planning/2010-demonstration-data-products/02-Demographic_and_Housing_Characteristics/2022-03-16_Summary_File/2022-03-16_Privacy-Loss_Budget_Allocations.pdf (accessed February 2023).

Van Ripper, D., T. Kugler, and S. Ruggles. 2020. "Disclosure avoidance in the Census Bureau's 2010 demonstration data product." In Privacy in Statistical Databases – PSD 2020, September 23–25, 2020, Tarragona, Catalonia. Springer: 353–368. DOI: https://doi.org/10.1007/978-3-030-57521-2 https://link.springer.com/chapter/10.1007/978-3-030-57521-2_25.

Y. Wang, Y., X. Wu, and D. Hu. 2016. "Usings randomized response for differential privacy preserving data collection." In Proceedings of the EDBT/ICDT 2016 Joint Conference, March 15–18, Bordeaux, France. DOI: https://doi.org/10.5441/002/edbt.2016.01; https://ceur-ws. org/Vol- 1558/article35.Pdf.

Warner, S.L. 1965. "Randomized response: a survey technique for eliminating evasive answer bias." *Journal of the American Statistical Association*, 60(309): 63–69. DOI: https://doi.org/10.1080/01621459.1965.10480775.

Wines, M. 2022. "The 2020 Census suggests that people live underwater." There's a reason. *The New York Times*, April 21. https://www.nytimes.com/2022/04/21/us/census-data-privacy-concerns.html (accessed February 2023).

Winkler, W. 1999. *The state of record linkage and current research problems*. Technical report, Statistical Research Division, U.S. Census Bureau. Available at: https://courses.cs.Washington.edu/courses/cse590q/04au/articles/Winkler99.pdf (accessed February 27, 2023).

Received February 2022
Revised August 2022
Accepted March 2023