Prof. Xiao-Li Meng
Editor, Harvard Data Science Review

Dear Xiao-Li,

At your invitation, we have submitted “The 2010 Census Confidentiality Protections Failed, Here's How and Why” (note the new, catchier title).

You asked us to address the critiques from the rejection at *Science*. This letter details the revisions that address those comments. In addition, this letter constitutes our permission to contact the *Science* editor (Ekeoma Uzogara <science_editors@aaas.org>) to obtain the names of those referees, provided of course that is within their policies.

Editor’s comments (in italics, replies in normal font):

We have now received the detailed reviews of your paper (attached). Unfortunately they are not positive enough to support publication of the paper in Science. Although we recognize that you could likely address many of these specific criticisms in a revised manuscript, the overall nature of the reviews is such that the paper would not be able to compete for our very limited space. We hope that you find the comments helpful and appreciate your patience during the review process.

We have addressed the reviewers’ comments. The editor fundamentally disagreed with Reviewer 3’s very positive review. That reviewer clearly thought the paper met the *Science* bar for importance and was well-written even in the abbreviated format that journal prefers. In response to the Editor’s comments, we have:

- Put the details that were in the appendices back into the main article;
- Developed graphical summaries for the main results and moved most detailed tables to the appendices;
- Extensively re-written the solution variability section because our abbreviated write-up was not as clear as it should have been about the conditioning sets we used;
- Extensively re-written the results section to make it as clear as possible that confidentiality breaches occur when the gain in predictive precision from using data that include the target individual vs. not using those data is large (it does not have to be infinite);
- Included some details of the extensive research on alternatives to the 2020 Disclosure Avoidance System in our discussion of how the DAS prevents reconstruction-abetted reidentification attacks.

Review 1 comments:

Given its potential high significance, I was very much looking forward to reviewing this manuscript. It is indeed, to the best of my knowledge, the first instance of a reconstruction attack in the real world, something theorized to be a risk a long time ago and the main justification for Differential Privacy.

We agree with this comment.

I was however, unfortunately, not able to evaluate the manuscript. The manuscript is indeed quite cryptic and excessively hard to understand, even for a specialist audience. This makes it difficult to *understand the results and, most importantly, their significance for the census and beyond (as claimed in the manuscript).*

We worked very hard to make the longer HDSR version accessible, but I will note that when I gave Ron Jarmin a copy to courtesy review, he said "you hit *Econometrica*, but I was hoping for *AER*." I was blunt with him, so I will be with you too. To avoid critiques like Reviewer 1, we have to include the details ("weeds") that technical journals insist on. For example, Algorithm 1, which is our fundamental matching technology, uses a left and right file. If the record in the left file is found in the right file, both are removed from the at-risk records, then the algorithm moves to the next record in the left file. This means that a record is at risk for a match until one is found, and never again. The most highly sited critique of our risk analysis is the paper by Ruggles and Van Riper (see references), which does this wrong—allowing a record in the right file to be reused over and over as a "match." That algorithm estimates the expected match rate of a random record in a random block, which is not what we are trying to do. Without the full details, the reader has to figure this out by themselves.

For instance, the authors dismiss SDL procedures altogether but seem to later suggest that they actually didn't evaluate the protection offered by SDL ("our attack did not try to undo the SDL"). The SDL procedure is furthermore, to the best of my knowledge, not described in the paper.

Frankly, I think this referee just didn't want to understand the SDL procedures we were undoing because they were clearly stated in the appendices to the *Science* submission. We relied on citations to describe the 2010 SDL in the short *Science* version. In this version, Section 2 describes them in detail, and Section 9 summarizes how they failed. Appendix C shows how we simulated them in our swapping studies. In addition, there is a complete summary of our attack in Section 5 that includes a footnote explaining how to extend the attack to other 2010 Census features. The explanation of what we mean by "not trying to undo the SDL" is in Sections 7 and 8, see fn. 29 as well. We when use the term "undo" we now always refer to the swapping, not the SDL in general. We do explain how to undo the swapping, but we did not actually undo it. We are undoing the consequences of aggregation.

Similarly, the paper lacks a clear description of the attack model(s) considered, a prerequisite to evaluate the results and their significance: what does the attacker have access to, what are they trying to do/infer, and why is this a concern.

See Section 5 "To make the meaning of this exercise as clear as possible, we state the attacker model we are simulating concisely here. ..."

There are, throughout the manuscript (a lot of) results and numbers but it's really hard to evaluate what they mean in practice and what the risk is.

Figures 3-4 and 5-6 address this critique. They show clearly that the reconstructed microdata have very high reidentification precision rates compared to the statistical baselines for nonmodal persons, especially in low-population blocks (Figures 3 and 4) and that this is not the case once the 2020 DAS is applied (Figures 5 and 6). Furthermore, Table 10 shows that the reidentification precision of the reconstructed microdata is essentially identical to that of the confidential data for the 3.4 million nonmodal population uniques who have 0 solution variability, whereas the statistical baselines for this group are essentially 0 (modal guesser) or the expected value of a random guess (proportional guesser).

The authors mention e.g. that "2010-era commercial data did not align well with the data collected in the 2010 Census" and seem to validate their inferences using a dataset held by the census bureau, something that would not be available to an attacker and might question the practical significance of (at least some) of the results.

Table 3 now clearly explains the problem with the 2010-era commercial data. Figures 3, 5 and Table 10 show that once this problem is addressed, the low-quality commercial data do enable confidentiality breaches with precision far in excess of the statistical baselines. Using the quasi-identifiers from the CEF itself is a reasonable approximation of high-quality name, address, sex and birthdate information. We should not have to actually produce an attacker dataset like that when it is well-known that commercial, redistricting, and school district databases do have such high-quality data now.

Later on, e.g., the authors also mention that only 3.4M people are actually population unique with zero variability.

The reviewer didn't read carefully. 3.4 million *nonmodal* people are population unique with zero solution variability. That's the most vulnerable population, and they were exposed by the 2010 SDL with precision in excess of 95% (note that since there is 0 solution variability, the 100% confidence interval on this estimate is +/- 0 percentage points). But trivializing 3.4M people is also inappropriate, especially since you can tell exactly who these people are.

Finally, the authors make very strong general claims on the ineffectiveness of SDL to prevent reconstruction attacks. The census is an extremely large data release: 50 billion aggregate statistics being released, something we know to strongly influence the feasibility of a reconstruction attack. It is thus not clear how generalizable the findings are to other data releases.

See footnote 19. We would prefer not to take down the American Community Survey or the Economic Census. They are vulnerable too, but we have not worked out all the mathematics. Every feature of the 2010 Census is vulnerable, as this footnote makes clear.

Reviewer 2 comments:

This paper presents findings from reconstruction and reidentification attacks the authors undertook for published microdata products from 2010 U.S. Census of Population was assesses the that assesses the confidentiality properties of these data and the Disclosure Avoidance System (DAS) used on to produce these products.

No, what we undertook for published *tabular* products was to show that they are equivalent to microdata. This referee is clearly familiar with the controversy surrounding the adoption of DP for the 2020 Census. His or her comments reflect many of the misinterpretations of what we did, which I will try to clear up here.

Some, if not most, of these findings were previously disseminated as part of the first author's Declarations as part of the State of Alabama v U.S. Department of Commerce (found at: https://www.census.gov/about/policies/foia/foia_library/frequently_requested_records.html) and in *a Census Bureau presentation (found at: https://www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series/ simulated-reconstruction-abetted-re-identification-attack-on-the-2010-census.html).*

Only a very small percentage of what appears in this article (or the *Science* submission) was exposed in either my Alabama and Fair Lines declarations or the ARSIA paper with M. Hawes. At that time the Census Bureau would not permit release of the full reconstruction/reidentification study (until 2022), and once approved, we tried to respect *Science's* pre-publication bar, which I independently confirmed with the editors. The complete study presented in the *HDSR* submission includes adding extra tables (SF1 P10 and P11), which increased 0 solution variability blocks by 5 percentage points.

With respect to the submitted paper, let me state at the outset that I found it extremely hard to read and evaluate. This is due, in large part, to the very technical way the authors wrote their paper. To some extent, their approach is understandable, given the technical nature of the design of the reconstruction and reidentification analyses they undertook and the somewhat arcane (in the broad sense of this term) nature of the structure the production of products from the decennial censuses of population. However, I don't think this can justify most of the difficulties that I had reading and evaluating this paper and the difficulties that I anticipate readers of Science will have in evaluating it.. Let me try to summarize the bases for this claim.

• The paper attempts to cover a number of somewhat extraneous issues that detract from what I see as the main contribution of the paper, namely, a detailed presentation of the design and findings from the Census Bureau's simulated reconstruction and reidentification attack on the publicly released 2010 Census data. For example, the authors include at the end of the paper ("Protecting the 2020 Census from This Attack") a discussion that attempts to compare its evaluation not only the DAS used in the 2010 public release products but also comparative analyses with the Bureau's DAS used for the 2020 Census based on differential privacy criterion. The issues surrounding the use of a differential privacy (DP) based DAS for the 2020 Census has been well-covered in other publications by various sets of the authors of this paper. The discussion provided at the end of the paper is an inadequate treatment of this comparison. This paper would be improved by omitting this discussion and attempting to publish a paper on this issue with the results from attacks conducted on the 2020 Census data directly (see lines 9-11 on p. 17).

We disagree and have had this discussion with our colleagues at the Census Bureau as well. There is no paper that shows how the choices for the 2020 Census addressed the risks exposed by our reconstruction and reidentification studies. We included Sections 9 and 10 and Appendices C and D so that readers could access all the data needed to see this for themselves. Yes, it makes the paper longer, but it also shows that protecting vulnerable populations is not a matter of just "turning up the swap rate" or doing some suppression. You have to have a workable definition of "vulnerable," which is what leave-one-out analysis provides. Then, you have to show that a comprehensive framework designed to protect all vulnerable populations works.

• The authors also attempt to answer critiques by others of their reconstruction and reidentification attacks. Again, this issue has been covered elsewhere (see Abowd Second Declaration found in the first above link) and more adequately there than in the discussion contained in this previous coverage.

We agree with this point. We only cite the critics of our analysis when we correct one of their mistakes. Otherwise, we leave to the supplement in Jarmin et al. the full brunt of the replies to those critics.

• While the authors may feel that addressing all of these issues in this paper were necessary, taking on all of these issues within the space constraints for an article published in Science from amount of published data is one of the reasons the paper is nearly impossible to evaluate. Too many of the key arguments are relegated to the Appendices in the Supplementary Materials and not adequately summarized in the main text of the paper. Thus, a reviewer and I fear readers are required to read through both the text and the appendices to even begin to follow the paper. My sense is this should not be required for readers; enough of the key arguments of the paper should be in the main text. In a sense, the authors are using the Supplementary Materials section of the paper to try to publish a 79 paper rather than a 2000 – 3000 word article as per Science article guidelines. I recognize that meeting the word count restrictions of a Science article are challenging, especially given all of the issues noted above that this paper attempts take on. But the submitted main text simply doesn't adequately cover all these topics without a fundamental reliance on also having to read the Supplementary Materials.

We also agree with this point, although we did submit under the 8,000 word limit for "online only" *Science* publications. The paper is much clearer if the reader does not have to constantly refer to the supplemental materials. We used *HDSR's* length flexibility to write a coherent main text. The appendices are genuinely supplemental information.

But my concerns about the current draft of this paper are deeper than what goes in the main text vs the Supplementary materials sections. There are a range of claims, i.e., claimed "unique contributions," for their paper made at the outset of the paper (see the Abstract and the opening bullet point paragraphs of the paper), that are simply not substantiated in the paper. Again, I provide some points in the paper where I found this lack of substantiation.

• In the Abstract and opening section of the paper, the authors argue that the reason they were able to reconstruct the confidential data from the publicly released 2010 Census products with what they argue is a high degree of success is the result of the fact that the methods used in the DAS for the 2010 did not satisfy the "mathematical property of composition. As they state in the Abstract, "The fatal flaw in the methods used for the 2010 Census is that they do not compose. (p. 1, lines 24-25). As the not in footnote ii on p. 21, "the mathematical property of composition means that if the methods are both used on different products from the same confidential input, the resulting publication still protect confidentiality—one method does not undo the protection of the other." While this property is an important issue in the development on the differential privacy criterion by Dwork and collaborators, I failed to find any evidence from the results reported in the main text of the paper or the Supplementary Materials that established that the failure of the 2010 DAS to compose accounts for the reported findings from there reconstruction attacks discussed in the main text or appendices in the Supplementary Materials. (In fact, this issue seems to receive no discussion after p. 3 of the main text.) There are other potential sources for the disclosiveness of the publicly released 2010 data. For example, that data was based on actual population counts by state and by finer levels of geography. These counts, or what are called "invariants" in the discussion of the DAS for the 2020 Census, can be informative in the reconstruction of the true underlying data

from the DAS protected released data. Furthermore, as noted by Gong and Meng (2020), the presence of these invariants may violate the requirements for mathematical property of composition. Without a direct assessment of the role of composition that accounts for features of the data like invariants, it seems inadequate to pin all of the disclosiveness of the 2010 Census DAS that the authors find on the former.

This is also a fair point, and we have taken pains to elaborate. We no longer refer to a failure of composition. Instead, we say that the 2010 SDL framework failed to defend against reconstruction, which turns tabular summaries into microdata that do not satisfy its own requirements for microdata releases. This is explained in Sections 2 and 9, and all claims about the failure of the 2010 SDL use this language.

• The claims that the paper contains "mathematical proofs" about the construction of "the exact image of the underlying confidential records for the stated feature set" (lines 24-26, p. 2) and of "an upper bound on the percentage of reconstructed records that can differ on no more than a single bit from their confidential image..." (lines 27-28, p. 2) are don't seem to be actually provided in either the paper or the Supplementary Materials. These may be established elsewhere, but I don't seem them adequately established, making it misleading to claim them as "unique contributions" of this paper.

The proofs of these claims are now clearly addressed in the methodology Sections 4 and 5, in particular in the re-written solution variability subsection 4.5.

• The discussion around the last paragraph on p. 14 and the follow-up discussion on pp. 14-15 concerning the production of reidentification using baselines based on MDG (modular guesser) and PRG (proportional guesser) reconstructed data that are similar to those based on the reconstruction methods used in the paper is confusing and inadequate. The authors claim that the similarity of reidentification results for the MDG and PRG reconstructed data files and those they present are misleading. They claim that the former violates their leave-one-out reasoning basis for inferences concerning what are legitimate inferences for identifying individuals in the data versus general "statistical inferences" about identification of individuals (or types of individuals) in the data. I simply don't follow this reasoning. A reidentification attack is primarily about identifying individuals that are in the data. As their findings indicate, they do not identify all individuals in the data, but rather disproportionately identify individuals with "unique" characteristics (sex, age and race/ethnicity) in less-populated Census blocks. Finally, as the authors note on lines 1-7 on p. 16, their reconstructed data files do better for population uniques than does the MDG and PRG reconstructed files. But what is so crucial about the leave-one out reasoning? This case is not made in the main text of the paper and not adequately in the Supplementary Materials. More generally, I don't understand why they claim that it is crucial to rely on this leave-one-out reasoning with respect to inferences. Even after reading their discussion of the leave-on-out reasoning for inferences about reidentification in the Supplementary Materials and I don't follow this claim. I would also note that relegating this discussion to the Supplementary Materials text seems inappropriate if indeed this is so crucial.

We have tried to write the entire *HDSR* submission to take on this comment because it is at the heart of the traditional SDL claim that as long as there is some uncertainty regarding whether or not a particular respondent's data were used the agency has provided "plausible deniability." We do agree that this did not belong in the Supplemental Materials, it should have been in the main text all along. What we establish is that the attacker does not need access to the confidential data to know (1) that the record is an exact copy of the confidential HDF (using the age-bin schema) and (2) that the record is a population unique on the quasi-identifiers (block, sex, age) used for a classic record linkage attack. While this is strictly true for “only” 97 million persons, adding more tables and more features can only increase this number. If the reconstructed microdata meet the conditions for a microdata-based record linkage attack, then such an attack must be defended in the SDL framework. So, the “plausible deniability” argument turns on whether the swapping was adequate. We argue in Sections 9 and 10 that it was not. In particular, if the released tabular data have reidentification precision that is essentially identical to the original confidential data then either those data could have been released without aggregation, which the Census Bureau has consistently argued would violate Title 13, or the tabular data are too disclosive, which is what we conclude. The burden of proof now lies with this reviewer to defend the status quo.

Let me return to my concern that the paper is difficult to follow. Here I provide some examples of such difficulties.

• The notation used to describe the various “feature sets” of the various data files makes reading the paper really challenging! I recognize the need to try to summarize these data files in a parsimonious fashion, but some of the later discussion of the paper could be vastly improved by providing readers with reminders in words as to the definitions of these data files! Relatedly, it would be extremely helpful to this reader if the authors had included abbreviated descriptions of these data files in Table 1. As it stands, one must constantly go back and find the definitions in the text of the “Data Used for the Experiments” section of the paper.

This is difficult to address completely. We made Table 1 more readable. We use consistent acronyms and feature names throughout the paper. That’s the best we can do without repeating definitions constantly. We do frequently remind the reader that *agebin* is the 38-bin age variable supported by the block-level tables.

• Table 3 includes a column “Block Population Range.” While I know that the authors are referring to the ranges of population sizes of Census blocks, this needs to be stated explicitly in the notes to this Table.

Done.

I also think authors need to include more of a discussion of the properties of Census blocks, since this is the unit of geography used throughout their analysis. Many readers won’t know much about these properties, such as their population sizes and the fact that blocks are not necessarily designed to have the same or similar population sizes. It would also be useful to refer to the geographic hierarchy of the 2010 Census. This could go in the Supplementary Materials or reference a discussion of it from the Census Bureau website.

We discuss this more, but we also provide a citation to the definitive census.gov explanations for readers who are unfamiliar with census blocks.

Finally, the paper fails to reference the crucial issue of the inherent tradeoff between protecting privacy and producing data that this accurate and usable. I recognize that this paper is about the *adequacy of the 2010 DAS for protecting privacy. And, implicitly, the findings concerning reconstruction do indicate that the 2010 Census data products appear to be too accurate. It they agree, I think the authors should make this point. This is especially important considering the controversy and criticisms of the adoption of a differentially private based DAS for the 2020 Census vis-à-vis concerns about the loss of accuracy of the data, especially at finer levels of geography.*

We do address this now in Section 10 and Appendices C and D. In addition, the replication archive contains the official metrics released for the 2010 demonstration DHC (and the link to census.gov) and the same metrics for the two swap experiments discussed in the paper. Appendix D deals with the accuracy of suppression.

Reviewer 3 comments:

The manuscript tackles an important issue on both technical and policy grounds. The issue is important from a technical perspective: there exists a long-standing literature on re-identification/de-anonymization, and a debate over how realistic and successful reidentification attacks can be "in the wild" (ie in real-life scenarios). In turn, that debates has implication for policy: depending on how accurate those attacks can be demonstrated to be in the real world (and therefore how concerned we should be about them from a public policy standpoint), initiatives such as the US Census adoption of differential privacy (which, in turn, has too attracted some degree of controversy) will be judged differently: valuable and important ways to deflect re-identification risks (if we believe those risks to be material and practical); unnecessary complications (if we believe those risks to be merely abstract and theoretical).

We agree with this comment.

In my assessment (which I completed with the assistance of an expert in re-identification), the manuscript provides both a significant technical contribution and a significant policy contribution. From a technical perspective, the results are novel and (to me) quite surprising: re-identification attacks using Census data are not just possible (this had been discussed before in the literature) but surprisingly accurate and precise. The significant policy implication arises from the fact that the manuscript also shows that the application of differential privacy can in fact protect from those attacks. Therefore, in a nutshell, the manuscript makes important contributions to two related streams of work: 1) the statistical re-identication debate and 2) the debate at the overlaps of statistics and public policy around the downstream practical implications of the deployment of DP or other privacy preserving analytics in census data, or other sensitive databases.

We agree with this comment.

Strengths

1. This is a striking and significant result, detailed thoroughly and rigorously. These results were clearly achieved with considerable effort, time, and expertise. They constitute a watershed finding for public statistics and data governance.

We agree with this comment.

2. The methods for reconstruction, reidentification, and analysis are very thorough and apparently sound. The Appendix is especially useful. Exceedingly transparent and detailed about precisely what was done.

Most of what was previously in the appendix is now in the main text with more detailed explanations.

3. The solution variability metric is a particularly clever device for expressing an attackers' confidence (using only the reconstruction results) and targeting the re-identification attack based on that confidence. The analysis of solution variability across blocks is helpful for understanding the mechanics of the reconstruction and leave-one-out thinking (in particular, that solution variability is very low for the small-population blocks where reidentification is easier).

We agree with this comment, and we rewrote the solution variability section to make the conditioning sets clear. Every claim is also proven.

4. The paper does well to set "scientific inference" baselines and clearly show how the attack exceeds those baselines for unique records in small blocks. This argument is cogent and clearly apparent in the results.

We agree with the importance of this argument. We hope that the clarified examples improve its cogency in this submission.

Issues

1. The language is very technical, and contains a lot of jargon and census-specific details that require several readings to understand. At best, this makes the paper hard to read for a general audience; at worst, there could be hidden idiosyncrasies in the many processes and convolutions involved with census operations that materially affect the results. As far as I can tell, this is not the case, but more clear and complete explanation would inspire more confidence. Some particular examples:

a. Fig. 1 is very helpful for understanding the relationship between all the acronymized datasets. It might also help to show here the relationship between HDF and CEF for completeness, and the relationship between HDF and SF1 (i.e., incorporate the latter half of Fig. S1 in Fig. 1). These relationships are key to understanding the empirical design, but as the figure appears now, it's unclear that SF1 is produced from HDF, which in turn is produced from applying SDL to CEF.

See the new notes to Figure 1.

b. Pg. 1, L23: The abstract mentions "precision", referring to a technical definition later in the paper. This definition requires some context — in particular, how a putative reidentification is defined, which depends on an assumption about the quality and content of attacker data. This should probably be included in the abstract — a more plain language expression (e.g. something like "... an attacker can, within blocks with perfect reconstruction accuracy, enhance an identified dataset of more than 614,000 (Table S8?) block, sex, and age records to include race and ethnicity with 90% precision..."). Otherwise, this metric is a bit incomprehensible — 90% precision at what? Affecting how many? Using what attacker data? Leaving the metric out would be better than including it without the right context.

We addressed this in the revised abstract.

The introduction (and conclusion as well) could draw out a few more key numbers — precision relative to COMRCL vs. CEF_atkr, the number of putative identifications in each condition, etc.

We think the new Section 10 addresses this comment.

c. Pg. 7, L8: "We harmonized the feature sets for [COMRCL]..." Does "harmonized" simply refer to the mapping described in this paragraph, or other steps? Were there any important differences in the constructs used in COMRCL, compared to census data?

We clarified this in the equivalent paragraph of this submission.

d. The intro (Pg. 2, L8-12) sets up the results as a study of the impacts of previous SDL methods, but this aspect is somewhat disguised in the presentation of the results. As far as I can tell from the census-specific terminology, the CEF represents the data before SDL, and the HDF represents the data after SDL. So measures of CEF-HDF agreement expressly represent the effects of SDL, the metrics rHDF-HDF describe the efficacy of reconstructing the post-SDL HDF from the post-SDL SF1, and the metrics rHDF-CEF describe the effectiveness of reconstructing the pre-SDL CEF from the post-SDL SF1.

This is almost correct. The CEF-HDF comparison only shows the effects of the record-level SDL. Aggregation must also be considered an SDL treatment here, otherwise, the HDF could be released. The HDF-rHDF comparison captures the aggregation effectiveness. We show that the aggregation was not effective, and the current submission is much clearer about this.

i. If this interpretation is correct, the results could make this clearer — for example, the authors could say directly that SDL reduces agreement, and remind readers that the reconstruction/re-identification metrics HDF-CEF are high relative to the pre-SDL records. Using "Pre-SDL" and "Post-SDL" instead of CEF and HDF could simplify this even further (also "2020 DAS", instead of MDF). If this interpretation is incorrect, then the framing in the intro is a bit misleading, and the authors should clarify exactly what treatments to the data the reconstruction is overcoming. The Conclusions section does this well.

We are much more careful about this in the current submission. There are some nuances that this reviewer missed. MDF is the output of the 2020 DAS, but rMDF is the reconstruction from the published SF1 tables constructed from MDF rather than HDF. Showing that MDF and rMDF both protect confidentiality establishes that the DAS defends against reconstruction-abetted reidentification attacks such that both the tabular summaries and the microdata used to create them can be released. We say this explicitly in Section 10.

ii. To truly know the effect of SDL, it would be nice to know the effectiveness of the attack on pre-SDL statistics (a "control" condition). As far as I can tell, these results do not communicate the counterfactual in which no SDL was applied—i.e. if the reconstruction was performed on a counterfactual summary file produced from the pre-SDL CEF (rCEFCEF). How should readers think about this counterfactual? Sure, the CEF-HDF agreement gives an upper bound on rHDF-CEF — but how effective might rCEF be? In other words, does the 2010 SDL have any effect at all on reconstruction and re-identification, or is it indistinguishable from an attack on pre-SDL statistics?

This would provide an important baseline for comparing the 2010 SDL methods (via rHDF-CEF) to the 2020 DAS (via rMDF-CEF).

This is an interesting suggestion that we did not implement. We think the reader will infer from Table 8 that the record-level SDL in the 2010 Census was very light-touch except in blocks with populations of 1-9. Doing the complete exercise with a version of SF1 tabulated directly from CEF instead of HDF would be interesting for census blocks with very low populations. We decided that this point didn't need to be made beyond Table 8. In addition, we would not have been able to release the 2010 CEF-based tables to our replication archive because they were never cleared for publication, and still wouldn't be.

e. The use of the term "vulnerable population" is a bit confusing. On first read, I thought this meant populations that are already vulnerable (e.g. marginalized) even before the possibility of re-identification. But as far as I can tell, "vulnerable" here refers to the population of nonmodal unique records. It would be helpful to state this more clearly the first time the term is used (pg. 2, L30), or to simply use a different term. In plain language, what kinds of people are most affected by the re-identification attack? E.g. are these nonmodal uniques mostly rural minorities? Similarly, who exactly are the "vulnerable" individuals, outside of this technical definition? (E.g. an analysis of the 13.2 million nonmodal records in <100 pop. blocks—pg. 15, L25—would be useful.)

We tried to do this throughout in this submission. In particular, the revised paragraphs at the start of Section 5 (before we leap into the algorithm definitions) are very explicit about the attack assumptions and the meaning of a vulnerable population.

2. The authors compute MDG and PRG only for the putative reidentifications already identified with rHDF (Pg. 10, L5-8). As far as I understand it, MDG and PRG represent how well an attacker might do with only scientific inference. But if the attacker has only scientific inference, how would they know to only target the putative reidentifications, and ignore the rest of the attacker data? Allowing attacker access to the putative reidentifications seems to allow them more than just a scientific inference. The authors could state this explicitly, and also compute MDG and PRG over the entire set of attacker data for comparison.

We agree that the attacker shouldn't be allowed to have the putative reidentifications "for free," however, we did not make clear in the *Science* submission that that ship had sailed. Specifically, releasing tables P12 and P14 meant that reconstruction of block, sex and agebin would always be perfect! We make that point clearly in this submission. Given that the quasi-identifiers (block, sex, agebin) can be perfectly reconstructed, starting MDG and PRG at the same putative reidentifications as HDF and rHDF is appropriate. We do make the point in Section 10 that the 2020 DAS also reduces the putative reidentification rate because it introduces noise in these quasi-identifiers.

3. Pg. 2 claims to offer mathematical proofs (e.g. a bound on solution variability), but I only see empirical results in the main paper and appendix. Perhaps I am missing something.

Proofs are now in the main text.

4. How much could attackers realistically bridge the large performance gap between the COMRCL data and the insider CEF extract, in particular for vulnerable subpopulations? More details about *the commercial databases would be useful. Do these databases effectively include the entire census population? If not, what kinds of people are missing from these databases (e.g. children)? What kinds of people are missing from the set of putative re-identifications? Might their data be accurately available elsewhere (e.g. in administrative data)?*

The discussion of Table 3 now explains that the problem with the c.2010 commercial data is that the quasi-identifiers (block, sex, agebin) don't match the CEF well. We cite the 2020 Census study of commercial and administrative data (Brown et al. 2023) to reinforce that this is no longer the case.

5. How well does the re-identification attack work with less or different attacker data? (E.g. inferring race/ethnicity with only block & age.)

We didn't do this because block, sex and agebin can be perfectly reconstructed, so it wasn't salient. We can think of many other ways to get the counterfactual world (e.g., use the 2005-2009 ACS block-group tables), but we chose to focus on other robustness checks.

6. Pg. 19, L33-37: "When the attacker has… thus defeating protection from aggregation." How does correctly reidentifying 75.5% of the records "defeat" protection from 2010 SDL? According to the arguments in this paper, protection is "defeated" according to leave-one-out logic — i.e. when re-identification precision exceeds scientific inference. But, as the authors acknowledge (pg. 14, L23), MDG does reidentify race/ethnicity with 76.6% precision (Table 4). According to this paper, the reconstruction attack only truly defeats scientific inference — and thereby protection from aggregation — for the "vulnerable" populations, right?

No. Given that we used the mode of race and ethnicity as the definition of nonvulnerable, the MDG is always going to be perfect on the modal subpopulation. But rHDFb,t, still gets them right as often as the HDF. They are still being "reidentified," but they would be reidentified in this sense from any reliable independent statistical information. We've included two detailed examples (smoking and Montana) to try to make this point. You don't want to prevent scientific inference, so you can't evaluate the effectiveness of the confidentiality protections without using a counterfactual world in which the scientific inference still holds but the reidentifying inference does not. We've tried again to make this clear with multiple examples in most sections of the revised submission. We have extensively rewritten the discussion of the experiments to ensure that we quantify success relative to leave-one-out inferences rather than absolute raw numbers.

7. Do the rMDF results align with the mechanism's formal guarantees? Are there any relevant theoretical bounds on precision for comparison?

Yes, see the results in Kifer et al. 2022.

8. Why doesn't Table S11 (nonmodal precision rates for rMDF over all blocks) include PRG and MDG, like Table 6 (only zero var blocks)? From the MDG and PRG numbers in Table 5 I'd guess that rMDF is also close to PRG all blocks, but I'd like to be able to see the comparison directly.

We hadn't computed those results. We have now. MDG and PRG appear everywhere. We don't repeat them in all the tables but the appendix tables contain every comparison.

Sincerely,

John Abowd

John M. Abowd