



UNITED STATES DEPARTMENT OF COMMERCE
U.S. Census Bureau
Washington, DC 20233-0001

February 26, 2020

Mr. Jeff Hardcastle
State Demographer
4600 Kietzke Lane
Building L, Suite 235
Reno, NV 89502

Dear Steering Committee Members:

Thank you for your letter on the U.S. Census Bureau's adoption of differential privacy to protect the confidentiality of respondent data for the 2020 Census. The Census Bureau places great value in the partnership and support provided by your networks, and we appreciate your collective commitment to helping the Census Bureau meet its dual mission of producing high quality statistics about the nation, while safeguarding the privacy of our respondents and the confidentiality of their data.

In your letter, you raised twenty-six questions about the Census Bureau's adoption of differential privacy and the implementation of the Disclosure Avoidance System. Enclosed, you will find our responses to your questions. I note that Census Bureau staff were available to discuss these and any related questions at the FSCPE Steering Committee meeting on February 12, 2020.

Sincerely,

A handwritten signature in blue ink that reads "John M. Abowd".

John M. Abowd, PhD
Associate Director and Chief Scientist
Research and Methodology

Enclosure

Questions Regarding the Proposed Disclosure Avoidance System

a) To date, what was the process used for input in making the decision to implement DAS?

Answer: The U.S. Census Bureau's Data Stewardship Executive Policy Committee (DSEP) relies on input from a variety of sources when making decisions about the adoption, implementation, and parameters of the Disclosure Avoidance System (DAS). These include internal subject matter experts, the Census Bureau's advisory committees (the National Advisory Committee and the Scientific Advisory Committee), the Committee on National Statistics of the National Academy of Sciences, academic experts and researchers, privacy advocates, professional associations, federal and state partners, and many others. We also solicited public comment through a July 2018 Federal Register notice. The Census Bureau has also conducted formal tribal consultations with American Indian and Alaska Native tribal leaders. Engagement with these and other stakeholders is ongoing. The Census Bureau will continue to solicit and consider feedback to improve the DAS throughout the coming year.

b) What inputs (testimony, research, outside experts, etc.) did the Bureau's Data Stewardship Executive Policy Committee/Disclosure Review Board use to decide the final optimal privacy-loss budget (trade-off between privacy loss budget and data accuracy)?

Answer: The final privacy-loss budget for the 2020 Census has not yet been determined by the Data Stewardship Executive Policy Committee. The value of $\epsilon=6.0$ was used for the release of the 2010 Demonstration Data Products, with $\epsilon=4.0$ allocated to the person tables and $\epsilon=2.0$ allocated to the household tables. These values were chosen after review of data presenting the impact on various demographic statistics computed for a range of different privacy-loss budgets.

c) Can you provide the planned scope of DAS project, the datasets and programs that will be affected by DAS (near and long-term), and the implementation schedule?

Answer: The DAS being developed for the 2020 Census of Population and Housing is being written specifically for the 2020 Census and cannot be directly applied to any other data product. However, the scientific and technical advances made during the development of the 2020 DAS will inform and enable the future development of formally private solutions for other Census Bureau products.

d) Does the Bureau have research that shows the 2010 Census file could be reconstructed / individuals identified without use of outside data files? It would be helpful for us to know if the reconstruction could be done without the use of a commercial dataset(s); and, if not, what commercial datasets the Census Bureau has used in reconstruction exercises.

Answer: The Census Bureau has now performed two partial reconstructions of the 2010 Census publications. Each reconstruction produces a microdata detail file containing the sex,

age, race, and ethnicity variables for each of the 308,745,538 enumerated individuals that made up the 2010 Census. No external data is required for the reconstruction attack. It uses only tract and block-level tables from the PL94-171 redistricting tables and the 2010 Summary File 1.

The reconstructed micro-data, which contain block, sex, age (in years) for every person enumerated in the 2010 Census, can then be matched against any file also containing block, sex and age to acquire names, and addresses for these same individuals. The attacker learns the exact race and ethnicity that was collected or imputed as part of the 2010 Census for specific individuals. The Census Bureau's internal re-identification experiments used a large database of commercial information that was acquired in the course of conducting the 2010 Census (the providers' names cannot be released because of the terms of the acquisition contract).

e) We understand that the Bureau is considering other formal privacy systems for public data products that include tables for detailed race and Hispanic origin tables, family/household tables that were included in Summary Files 1 and 2 in 2010, an American Indian and Alaska Native Summary File, and the Public Use Microdata Sample (PUMS) File. What are the other formal privacy systems under consideration for these data products? When will information about these systems and the availability of products be released and what will be the opportunity for input by the networks?

Answer: There are many different ways of implementing formal privacy, and the optimal selection and design of these systems depends on the characteristics of the data and the specific use cases for which data accuracy is to be optimized. The TopDown Algorithm (TDA) central to the DAS was designed specifically to produce the PL94-171 (redistricting data), the Demographic Profiles, and the demographic and housing characteristics data files. For other 2020 Census data products, including tables for detailed race and Hispanic origin, and tables previously included in the American Indian and Alaska Native (AIAN) Summary File, and tables requiring person/household joins, we are planning a secondary system that will be added as an extension of the DAS but still based on differential privacy. Because of their level of detail, these tables pose difficult and unique privacy challenges. The Census Bureau remains committed to producing data on detailed race and ethnicity, AIAN tribal affiliations, and characteristics of people within households to meet our data users' needs. Contracts have been executed to implement differentially private methods to produce these tables and to modify the dissemination system so that they can be released, once produced. We anticipate being able to provide more information about these efforts in the coming months, at which point we will work with the data user community to ensure these products meet their needs. At the present time, no final decisions have been made regarding the release of a Public Use Microdata Sample (PUMS) file for the 2020 Census.

f) How is the Bureau coordinating the implementation of DAS across divisions and branches? What are the contingency plans if DAS cannot be implemented as currently envisioned?

Answer: The DAS is being implemented by an agency-wide, interdisciplinary team that is headed by the Chief Scientist.

The DAS is currently operational and able to perform the disclosure avoidance necessary for the legislatively mandated publications of the decennial census. As we work to improve and optimize the DAS for an array of priority data use cases, we are also researching a variety of contingency plans to ensure that the 2020 Census Data Products meet the Census Bureau's data quality standards.

g) Can you provide specific information regarding the criteria that will determine suppression by geographic level? We are concerned that data may not be available for key geographies including but not limited to blocks, block groups, tracts, census designated places, and minor civil divisions. It is of grave concern that DAS would have a substantial adverse impact on the availability and quality of data for small communities.

Answer: Suppression is a traditional disclosure avoidance technique that protects privacy by redacting or not publishing data for small groups or small geographies. One of the advantages of differential privacy is that noise infusion and the privacy guarantee removes the need for suppression by geographic level. Because of the impact of differential privacy on data accuracy for small geographies or populations, however, the Census Bureau is evaluating what tables to release and at what geographic levels to ensure that our data products meet fitness-for-use standards.

h) Can you provide similar information for other programs (American Community Survey (ACS), Economic Census, National Center for Health Statistics, etc.) on the tables or products that will be either modified or suppressed by using DAS or similar techniques, and the impact on geographies by FIPS and by NAICS for business data?

Answer: The application of differential privacy to the ACS, the Economic Census, and other data products is still in its research phase.

i) Will DAS also be implemented on data products produced for other agencies? Have these other agencies (such as BEA, the Department of Housing and Urban Development, the Bureau of Labor Statistics, the National Center for Education Statistics, and the National Science Foundation) and their data users provided comments? If so, in what form and where can the comments be accessed?

Answer: The 2020 DAS is custom developed for the 2020 Census and cannot be applied directly to multi-stage probability sample surveys, the bulk of the agency's contractual and internal products. More generally, the Census Bureau has indicated that it is committed to moving to formal privacy protection techniques for all of its data products, but the timetable for that

transition is unclear at present. We are committed to a vigorous and open discussion of the new DAS, and its impact on data quality and fitness-for-use. It has always been important for data users to understand the consequences of disclosure limitation procedures, and we know that these conversations will benefit the Census Bureau and improve our data products.

j) Will DAS be implemented on data products produced as a special tabulation? Has there been any research by the Bureau on special tabulation uses and the impact DAS has on these tabulations?

Answer: The mechanism by which formal privacy will be applied to special tabulations based on the 2020 Census is currently under development.

k) What criteria were used to decide on the proposed suite of the demographic and housing characteristic file (DHC) data tables?

Answer: To create the proposed suite of the demographic and housing characteristic file (DHC) tables, the Census Bureau started examining the suite of demographic and housing characteristics tables that were present in the 2010 Census SF1 and other data products and then began removing statistics that did not have a well-defined use case or that could not be readily computed using the 2020 DAS TDA. As we discuss in our response to question (e), we are developing alternate formally private systems to produce data products that cannot be computed using the TDA. Decisions have been made based on demonstrated use cases received through the July 2018 Federal Register Notice and through extensive stakeholder engagement and outreach by Census Bureau subject matter experts. If tables at particular geographic levels did not have a demonstrated use case, they were considered for removal from the DHC. This stakeholder outreach and engagement is ongoing, and we will continue to revise the proposed suite of data products as we receive additional information from our data users.

l) Has the Census Bureau considered expanding the privacy budget as the 2020 Census data becomes less current and therefore less valuable to potential reconstructors? Typically the Census Bureau does not issue Summary File 2 data until 2 or more years after the census is taken. Is there consideration in balancing privacy, currency, and accuracy in the implementation of the DAS?

Answer: The Census Bureau is currently engaging with a variety of stakeholders on this and related questions. No decisions have been made at this time.

m) For the tables that will be released for which DAS has been applied, what impact will it have on the ability to compare tables over time?

Answer: We recognize that the adoption of differential privacy for the 2020 Census will have implications on time-series and trend analysis when using the 2020 data products in

conjunction with data from prior censuses and surveys, and we will work with our data users to provide guidance on how to compensate for these effects. Each table in the DHC will be released a single time in 2021. It is premature to discuss the comparison of the 2020 DHC with either future versions of the ACS or with the tables that will be released for the 2030 Census.

n) Has the Bureau researched the impact of not producing data that was previously available? Has the Bureau considered that federal, state, and local programs in many cases, have statutory or administrative mandates requiring the use of Census data for funding and for reporting? Many of these requirements have been developed because accurate Census data provides a way to equitably distribute state and local revenues and services.

Answer: The Census Bureau is committed to continuing to produce the high quality data on which our many stakeholders have come to rely, consistent with our statutory obligations to protect the privacy of our respondents and the confidentiality of their data. To ensure that we are meeting our data users' needs, we have already engaged in extensive outreach to the data user community to catalog the various statutory and administrative uses of decennial census data. On July 19, 2018, the Census Bureau solicited "feedback from users on 2020 Data Products" (83 FR 34111). A second solicitation (83 FR 50636) appeared on October 9, 2018. The Census Bureau also participated in a meeting of the Committee for National Statistics on December 11-12, 2019, where major data users presented the results of their analyses of the 2010 Demonstration Data Products. More generally, the Census Bureau is eager to engage with federal, state and local programs to learn more of how they use census data and their requirements for accuracy. The Census Bureau is also eager to engage with stakeholders to understand the privacy expectations, requirements, and concerns of the American public. The Census Bureau's ultimate objective in engaging with these stakeholder groups is the development of principled approaches for balancing the needs and statutory requirements of both communities.

o) Most Census survey data already have margins of error (MOE). Does applying DAS compound these errors? The three networks already hear many complaints about MOEs – particularly for small geographies. We are concerned that, not only will the data be less usable, but that Census survey respondents will be unwilling to fill out forms if they perceive the resulting data is less accurate and unhelpful for their needs.

Answer: Error in census and survey statistics comes from a variety of sources (coverage error, measurement error, etc.), and privacy protections of all types add to this overall error. One major advantage of differential privacy over traditional disclosure avoidance techniques is that it allows the error (uncertainty) resulting from disclosure avoidance to be measured, reported, and discussed in a fully transparent manner. In general, the Census Bureau expects that the impact of the error introduced by the use of formal privacy will be less than the error resulting from other factors.

p) Why did the Bureau take the proactive stance to be the global leader in implementing disclosure avoidance, when so many vendors are collecting, selling, and publishing data that are not under the Bureau's control? It is important that our network members, data users, and the stakeholders we serve understand why the Bureau took the action proactively to be the global leader in disclosure avoidance without, as far as we know, any major challenge that the Bureau was not upholding its 13 USC mandate. It is also important to acknowledge that the Bureau's initiative will not solve the global problem of personal data disclosure.

Answer: The Census Bureau has been a global leader in the design and implementation of disclosure avoidance methods for decades. The Census Bureau's decision to adopt differential privacy is merely the next step in a history of innovation of our privacy protection methods to counter evolving privacy threats. The fact that there are many vendors collecting, selling and publishing data about United States residents does not lessen the Census Bureau's obligations under Title 13. Furthermore, although the coverage of commercial vendors is very good for some segments of the population, it is quite poor for others. Specifically, commercial data do not have good coverage of children, of self-reported race, of the existence of same-sex relationships, or of parents who have a different race than their children. This is precisely the sort of information that is collected by the decennial census. Given that the decennial census is a mandatory survey with universal coverage, the Census Bureau believes that it has both a legal and an ethical responsibility to use the strongest privacy protection technology available. The Census Bureau has a dual mandate to produce quality statistical information while protecting the confidentiality of respondent data. We know that the nation needs timely, accurate information to make informed decisions. Our goal is to ensure that the public trusts us with their data and values the statistics that we produce. Adopting our advanced confidentiality protection system helps us to meet that goal.

q) What other methods or consequences did the Bureau consider for protecting privacy, either legal or methodological, which would fulfill the Bureau's duty to protect an individual's record(s) and still produce data that can be used by the everyday user and local elected officials? How is this being handled by other governmental statistical agencies both within and outside of the US?

Answer: We know of no other statistical technique that can be reliably employed to assure the confidentiality of the underlying data while simultaneously assuring the highest quality statistical product for our data users. Other privacy mechanisms, such as the k-anonymity technique or even the swapping technique that was used in the 2010 Census, are now generally recognized by privacy researchers as being insufficient to meet twenty-first century privacy threats. In this era of Big Data, simply adding more noise using our older methods is not a workable solution. So much noise would be required that our published data would be unfit for most uses.

r) Has the Bureau considered the consequences of the implementation of DAS on non-governmental entities and programs that provide key community services? Has the Bureau presented the impacts and received input from small states, local, and non-profits on the implementation of differential privacy? A major concern is whether these organizations will have usable data to conduct research, present a case for grant funding, and build the right sized and type programs needed for their communities.

Answer: The Census Bureau is committed to publishing accurate data for the 2020 Census, however our obligations to protect privacy mean that we cannot publish perfectly accurate data for every conceivable use case. Based on the stakeholder feedback we have, and continue to, receive, we are endeavoring to ensure that the 2020 data products meet as many of our data users' needs as possible. The Census Bureau is also committed to maintaining the scientific integrity of the analyses performed using the public-use products that the Census Bureau releases. The Census Bureau will issue suitability for use guidelines that reflect the effects of the DAS. We will also publish our final algorithms, and the parameter values used by those algorithms, so that researchers can use the data in a scientifically appropriate manner. To conduct scientific analyses for which the public-use data are not suitable or sufficiently accurate, researchers may choose to seek approval to conduct a project under the auspices of the Federal Statistical Research Data Centers.

s) Has the Bureau determined the impact on program reports whose findings may be distorted due to the implementation of DAS and may no longer accurately represent the reported geographic area, population group, or economic sector?

Answer: Such work is currently ongoing. The Census Bureau is eager to work with stakeholders to develop systems such that the impact of the DAS on geographic areas, population groups, and economic sectors can be quantified and minimized.

t) Has the Bureau considered that the implementation of DAS will result in limited data availability for small geographies, leading these entities or service providers to purchase data or conduct surveys through private companies? The profusion of companies willing to provide data and surveys may compound disclosure issues since they are not subject to 13 USC requirements and will not use the same strict methods and guidelines the Bureau employs for both data collection and tabulation. For example, with the lack of state level population projections the private sector has stepped in from a variety of vendors with different products and levels of transparency.

Answer: Although the data we produce for the 2020 Census will be infused with noise, the 2020 DAS is designed such that statistics computed on larger populations, such as block groups or census tracts, will be significantly more accurate than statistics computed at the level of a single block. We agree that the proliferation of third-party data sources poses serious privacy concerns for our respondents. Since the last decennial census, the data world has changed dramatically. Growth in computing power, advances in mathematics, and easy access to large, public databases pose a significant threat to confidentiality. These forces have made it possible

for sophisticated users to ferret out common data points between databases using only our published statistics. If left unchecked, those users might be able to stitch together these common threads to identify the people behind the statistics. Because we are sworn by law to protect our respondents' data, we are constantly testing and improving our privacy protection methods to stay ahead of these changes. Our adoption of differential privacy for the 2020 Census is necessary to ensure that as more of these third-party data sources emerge and improve over time, they will not weaken or erode the privacy guarantees we provide to our respondents.

u) What plan does the Bureau have to inform and work with data users to ensure the implementation and impacts of DAS do not have negative consequences?

Answer: The Census Bureau is actively working to better inform and engage with our data users and the broader American public regarding all of the Census Bureau's efforts to protect respondent privacy while providing high-quality statistics about the nation. These ongoing efforts have taken many forms, including the December 2019 workshop sponsored by the National Academies' Committee on National Statistics. While many data users might wish for data to be published as accurately as possible—that is, without any privacy protection—there are also many respondents for whom privacy is a major concern, making the Census Bureau's ability to safeguard respondent data vital to the Census Bureau's efforts to maximize response rates. In the end, all statistical projects and efforts to protect respondent privacy have both positive and negative consequences that must be balanced by policy makers.

v) After data release, how is the Census Bureau going to handle criticism from the public that starts questioning the quality of the data because they find implausible numbers or don't recognize themselves or the area they live in in the published numbers? Will our networks get any guidance on how to deal with that criticism?

Answer: The Census Bureau is actively working to improve the 2020 DAS so that there will be few such implausible numbers in the 2020 data products.

w) What is the process, format, and timeline for the three networks to provide input to the Bureau? The input would include both the specific impact of DAS on the data for governmental and non-governmental organizations, as well as the result of network member comparative analyses of the demonstration tables and the 2010 tables.

Answer: Improvements and optimization of the DAS are ongoing, and will continue throughout most of calendar year 2020. Consequently, we welcome any data user feedback on the 2010 Demonstration Data Products that we may receive through the summer of 2020. While feedback received earlier in this process will have the greatest potential for informing major changes to the DAS design and configuration, we will assess and consider all feedback that we

receive on the Demonstration Products on an ongoing basis until the configuration of the DAS is finalized in late 2020. Partners may submit their feedback on the Demonstration Products individually or collectively by submitting them to dcmd.2010.demonstration.data.products@census.gov.

x) The Census Bureau has asked each of the partnerships to provide support for the implementation of differential privacy. Can you please provide what you are requesting each partnership to do to show support?

Answer: From our perspective, the most helpful feedback we could receive from the partnerships, in addition to the obvious identification of impossible or improbable outcomes in the 2010 Demonstration Data Products, would be suggestions that could be used to improve the design and optimization of the DAS to produce data products with the highest fitness-for-use.

With the understanding that there are basic tradeoffs between accuracy and privacy that DSEP will need to navigate, the most actionable suggestions we could receive from the partnerships would include results-oriented objectives (e.g., “willingness to sacrifice some existing accuracy at the block level to improve tract-level data”) or standards-based thresholds (e.g., “county/tract/block-level data needs to be at least X/Y/Z percent accurate to be acceptable”).

y) How does injecting noise into the data, disconnecting household relationship and effectively changing population counts for small areas impact the Census Bureau’s residence rules and how local and state governments review the accuracy of the Census?

Answer: In past censuses, the Count Question Resolution (CQR) program has provided jurisdictions with the opportunity to verify the correct geolocation of group quarters facilities and housing units in Census Bureau tabulations. While the Census Bureau has not yet finalized details of how the CQR program will operate for the 2020 Census, we recognize that the operation of this program may be impacted by the transition to differential privacy. As these details are finalized we will engage with the partnerships to better answer this question.

z) How does the Bureau justify shrinking the availability of data about communities ranging from Asian ethnic groups and the Middle Eastern community when they have been asking for an expansion of how their specific communities are reported at least at a national level?

Answer: The Census Bureau is committed to producing data on detailed race, ethnicity, and American Indian and Alaska Native tribal affiliation at various geographic levels to meet our data users’ needs. As these products will be produced through a different formally private system separate from the DAS, they are not currently included in the 2010 Demonstration Data Products. As we continue with the design and development of this second system over the coming months, we will actively engage with the partnerships to evaluate and improve these additional data products.