Draft

Disclosure Avoidance and the 2020 Census: What Do Researchers Need to Know?

Daniel L. Goroff
Alfred P. Sloan Foundation
goroff@sloan.org

Erica L. Groshen
Cornell University--ILR School and W.E. Upjohn Institute for Employment Research
erica.groshen@gmail.com

Abstract

The Census Bureau's plans for the 2020 Decennial will publicly and transparently address the unavoidable trade-off between data privacy and data accuracy. Statistical analysts can, and should, therefore take into account the planned presence of well-specified and well-justified noise in data releases based on the 2020 Decennial Census.

To aid researchers' preparations, this paper examines strategies, trade-offs, and rationales associated with processing and releasing the Decennial results. We highlight both what is new as well as and what seems new but is actually little changed. Based on this review, we offer specific conclusions to help promote appropriate and well-informed usage of the 2020 Census. Our strongest recommendation is that, in addition to publishing official tables, the Census Bureau also make the Noisy Measurement File (NMF) available for research purposes. The latter dataset is privacy protected, but the negative or fractional entries that appear as a result may be more easily interpreted by trained analysts. Restoring the face validity to the counts as expected by other users will, in contrast, introduce statistical bias and intractable distortions that academic researchers may wish to avoid whenever possible. Our other recommendations address critical needs for input to Census decisions from researchers, for development of suitable statistical tools for privacy-protected data, for expanding options with regard to microdata, and for steps to improve the accuracy of Decennial Census data overall.

Disclaimer: The views presented here represent those of the authors alone and not those of any organizations with which they are affiliated.

Acknowledgements: The authors are grateful for helpful conversations, comments, and corrections from three anonymous referees, danah boyd, Katharine Abraham, Joseph Hotz, Mark Schweitzer, Michael Hawes, Gary King, Cynthia Dwork, Simson Garfinkel, Robert Sienkiewicz, John Abowd, Xiao-Li Meng, Ruobin Gong, members of the Census Disclosure Avoidance Systems Expert Group convened by the National Academies Committee on National Statistics, and participants at the October 25, 2019 Harvard Data Science Initiative Conference on "Balancing Data Quality and Disclosure Risk Using Differential Privacy Methods: Impact on the Decennial Census."

Disclosure Avoidance and the 2020 Census: What Do Researchers Need to Know?

"Sometimes, the riskiest thing to do is nothing." Timothy Geithner, 2007

I. Introduction and Summary

Among the many users of 2020 Decennial Census data, what should researchers, particularly, know about new disclosure avoidance procedures for protecting individuals' privacy? For data stewards and data users alike, there is much to learn. That is because privacy protection methods based on the idea of "Differential Privacy" provide provable and understandable guarantees in principle while posing difficult but unavoidable trade-offs in practice.

Specifically, 2020 will be the first Decennial Census designed to address—publicly, explicitly and transparently—the relationship between data privacy and data accuracy. The inevitability of zero-sum trade-offs between these two fundamental goals is a law of nature rather than a challenge that can be eliminated with better technology or procedures. Most commercial interests want us to believe that our only choice, if we wish to reap the benefits of "big data," is to sacrifice our privacy altogether. In contrast, federal officials seek to demonstrate how it is possible to strike a balance that respects the Census Bureau's legal obligation to provide both needed privacy protections *and* useful public statistics.

Despite much popular discussion about protecting confidential information, the only known conceptual framework for defining, analyzing, and implementing rigorous privacy protections is the theory of "Differential Privacy." The 2020 Decennial will not be the first time that the Census Bureau has employed disclosure avoidance procedures based on these ideas. Data products such as OnTheMap, Post-Secondary Employment Outcomes, and Veterans Employment Outcomes are previous examples. In each case, answers to statistical questions are infused with just enough carefully-calibrated noise to provably protect privacy while still providing useful results. Leading tech companies, including Apple, Microsoft, and Google, have also begun employing Differential Privacy. Previous public and private applications like these have been welcomed by data users because they allow access to sensitive information that simply would not have been safely accessible to independent researchers unless altered first by the addition of some noise.

Researchers are understandably more perplexed and skeptical about the role of Differential Privacy in releasing results from the 2020 Census. After all, everyone has been accustomed to receiving pages and pages of tabular data every ten years. It was easy to take the entries in those tables at face value. Even while pretending otherwise in this way, social scientists always deal with measurements that, whether privacy protected or not, contain errors and noise. Previous Disclosure Avoidance Systems introduced such inaccuracies, too, but were ad hoc, obscure, and largely undocumented. So the distortions those systems produced were unknowable in extent or magnitude. Nevertheless, they were conveniently assumed to be both

DRAFT

protective of privacy and small compared with undercounting or other systematic sampling challenges. We now know that these previous presumptions and procedures were not nearly as reliable as commonly assumed (see Section IV below). Unable to ignore such findings, the Bureau will actually produce three different kinds of statistics based on the 2020 Decennial:

S1. **Noisy Measurements:** These are counts that, after they are subject to imputation, de-duplication, and other traditional forms of pre-processing, have random noise added to them drawn from a probability distribution whose formula and parameters are set by officials and announced publicly. Depending on their choices, this mechanism provides an explicit degree of privacy protection in exchange for an explicit degree of count distortion (see Section VI below). In stark contrast to previous Disclosure Avoidance Systems, the theory of Differential Privacy also provides theorems that—when they apply—describe how these protections and distortions behave as further information is released, aggregated, or processed (see the Appendix below for more details). Note that the Census Bureau has not yet announced if it will release files with noisy measurements to researchers or the public.

S2. **Invariant Statistics:** These are counts that, after they are subject to imputation, de-duplication, and other traditional forms of pre-processing, are released directly without the addition of noise or any other form of distortion. The Census Bureau has announced that there will be very few such numbers published by the 2020 Decennial. Chief among them are the state-wide total population counts as enumerated. The reason, of course, is to bolster public confidence in the re-apportionment of Congressional seats based on those figures. Releasing invariants does prevent the straightforward application of basic theorems about privacy guarantees, but there are other interpretations, modifications, and approaches mentioned below that support the use of Differential Privacy even so.

S3. **Post-Processed Tables:** The noisy measurements mentioned above will include "counts" that are negative, fractional, or otherwise unacceptable as entries in published Census tables. The Bureau therefore plans to make adjustments that, while trying to stay as close as possible to the noisy measurements, will nevertheless produce published tables whose entries look appropriate (that is, have the same logical properties as tables based on the confidential data) and tally properly. This is an enormous computational challenge. It is accomplished using proprietary optimization software whose inscrutable calculations introduce statistical bias and other data distortions with properties that are hard if not impossible to characterize (see Section VII below). Standard theorems assert that "post-processing" like this cannot erode the confidentiality guaranteed by Differential Privacy—as long as the computations refer only to the noisy measurements and not to the unprotected data. The "TopDown Algorithm" that the Bureau will use to create tables (see Section VIII below) does make use of

DRAFT

invariant state population totals, however. That makes the implications of all this post-processing even harder to specify in detail.

This article will explain just enough about the processing of these three kinds of statistics to help academic researchers and other Census data users begin preparing for how 2020 Decennial releases will be different due to new disclosure avoidance procedures (for even more technical details, see other contributions to this volume). Based on our account of these changes, we conclude by making a number of recommendations. Our main message is that Census should release the Noisy Measurement Files (NMF) and that researchers should consider bypassing the official tables, despite their great importance for legal and other matters, and deal with the NMF instead whenever possible. What the NMF lack in face validity, they make up for in improved analyzability. Every empirical social scientist should have training in how to take random error terms into account. As noted, 2020 Census data, even the confidential, cleaned microdata, will have some errors of unknown structure. The difference here is that, unlike previous Censuses, where little can be determined about the Bureau's process for adding privacy-protecting noise, in this case the Disclosure Avoidance System (DAS) will add noise drawn from fully-disclosed distributions with fully-disclosed parameters using fully-disclosed code. Drawing valid statistical inferences is never easy, of course, and there are special challenges and solutions even when dealing with data like that in the NMF (see Evans et al. 2019, Evans and King 2020). But the post-processed published tables, in contrast, will include influences from opaque computations whose full implications for statistics are nearly impossible to gauge.

In what follows, we present some of the Census Bureau's rationale for its particular implementation of Differential Privacy—only to the extent that can help academic researchers and other Decennial data users better understand the new Disclosure Avoidance System. Our aim is neither to justify nor to pass judgement on the Bureau's decisions. Nor do we dwell on alternative approaches, unless these still have some chance of playing a role in how the data can be used and interpreted.

We do, however, want to acknowledge that the Census Bureau has faced up to new and difficult trade-offs. By all accounts, officials have focused seriously on honoring their dual obligations concerning privacy and accuracy, subject only to constraints imposed either by federal laws or by mathematical laws. And, in contrast to those implementing Differential Privacy for commercial purposes, they face much greater timing and scale challenges, in addition to higher obligations for transparency, uniformity, and accountability.

Not everyone will be happy about the new Disclosure Avoidance System, of course. The disputes and disappointments are not merely over algorithm design or parameter tuning. On the contrary, trying to deal soundly and forthrightly with privacy/accuracy trade-offs is forcing a fundamental reckoning with what Census data are supposed to mean and do. It is not an exaggeration to call this an epistemological crisis that, even within the academic research community, different interests approach quite differently.

DRAFT

Take, for example, modelers like economists and other routine users of regression analysis. While they may find the "sensitivity" of regression coefficients to outliers challenging when applying Differential Privacy techniques, they are hardly perturbed by the addition of random noise terms generally. But demographers and redistricting analysts, whose methods sometimes have more in common with those of accountants, naturally find systematic obfuscation of any sort unacceptable. Cryptographers and privacy researchers admirably begin by worrying about how to protect all forms of data from all forms of attack, including privacy threats that might seem unlikely to plague Census results. And statisticians, accustomed as they are to dealing with sampled data, fret over how the planned post-processing may render Decennial tables "uncongenial" if not immune to analysis in terms of a consistent data generating process.

These are all valid and respectable perspectives, as are urgent public concerns over confidentiality, fairness, legal matters, etc. Our hope in presenting this introduction to the Disclosure Avoidance System for the 2020 Decennial Census is to provide shared vocabulary, translatable concepts, and common ground that can help researchers of all sorts put the data to good use.

To summarize what follows, here is the basic story. Census takers work hard to count everyone, but there are always errors, noise, sampling biases, and other inaccuracies in the records they assemble, and especially in 2020 due to COVID and other fears. The Census Bureau routinely does what it can to correct for, or at least estimate, some of these deficiencies. But perhaps the biggest threat to the accuracy of any Decennial Census concerns response rates. Data about residents who fear or mistrust the government may be incomplete, falsified, or missing altogether. For that reason, federal law prohibits the Bureau from releasing information that makes individuals identifiable. Having found the *ad hoc* privacy protections previously used to be no longer credible, the Census Bureau is adopting a new Disclosure Avoidance System to release the 2020 Decennial. The first part of that system is based on the only formal, rigorous, and comprehensive framework available for protecting privacy, the theory of "Differential Privacy." Theorems and experience recognized since the last Decennial show that protecting privacy requires releasing each count only after adding some specific and specified noise. The probability distribution of that noise is public and has parameters, notably one referred to as epsilon, that officials can set and announce to adjust the trade-off between protecting privacy (more noise) and protecting accuracy (less noise). No one should therefore be surprised when such a random variable occasionally takes on a large or negative value. Nor should it be surprising, as a mathematical matter, that ratios (including percentages) or other nonlinear functions of privacy-protected counts come out different from past or expected unprotected values. That is how Disclosure Avoidance Systems work. There are challenges, as always, to drawing statistical inferences from noisy measurements. But we argue that researchers can deal with those challenges and should deal with those files. Census tables, on the other hand, have many important uses in realms other than the social sciences. To meet the public need for tables with face-validity, including non-negative entries and consistent tallies, the Bureau will engage in extensive and computationally intensive "post-processing."

DRAFT

Researchers should know that this will curtail their ability to conduct valid statistical analysis of probability distributions and other inferences based on those published tables.

II. Significance

Enumerating U.S. residents is a fundamental responsibility of government enshrined in Article 1 of the U.S. Constitution. The Decennial Census determines how seats in Congress are reapportioned among the various states. Many states also rely on the results to redraw Congressional and other voting districts, even when not required to do so by law. In fact, population tables released by the Census Bureau underlie nearly all the demographic descriptions and decisions made by government, business, or other organizations in the United States. Massive federal expenditures—including funds for Medicaid, Medicare, Supplemental Nutrition Assistance, Highways, and School Lunches—are distributed according to population estimates based on Census data, for example. Considering the 16 largest such programs alone, a study called "Counting for Dollars" tallied up nearly $1.5 trillion allocated this way in 2020 (Reamer 2020). So in contrast to elections, where a single vote can seem unlikely to make a difference, each person's decision about whether to follow the law that requires participation in the Census could send many thousands of dollars per year from one jurisdiction to another.

Researchers are active and influential users of Decennial Census data products. Just one access point, the Integrated Public Use Microdata Series (IPUMS) lists almost 2,000 citations, covering widely varied topics, that refer specifically to using the Decennial Census demographic and housing data files. (See the IPUMS search mechanism at https://bibliography.ipums.org/citations/search. Search for "NHGIS.") These represent only part of the over 12,000 research products about the US listed in the IPUMS system, most of which use Decennial Census data in some way. Recent studies of placed-based policies (such as enterprise zones) are particularly notable in this regard. Along with explicit uses like these, many other studies rely on Census data obtained from other sources and use Census data for somewhat indirect purposes that range from sampling design to validating, weighting, modeling, or augmenting more current or specialized sources. By drawing inferences from population estimates based on the Decennial Census, both academic and nonacademic analysts produce work that informs key policy and business decisions. And as statistical agencies move to apply formal privacy protections to other data releases, familiarity with the Census Bureau's current Disclosure Avoidance System may prove useful well beyond the 2020 Decennial.

What is new? Social media platforms now offer particularly diabolical ways to skew or misuse Census data. Disclosure avoidance procedures once considered adequate before personal data was so accessible over the internet now leave people open, both provably and practically, to the kind of re-identification that the Census is forbidden to allow by law. Such misuse can harm targeted groups. Even if not intended to deter future Census participation, it could do so as vulnerable groups try to avoid the risks of harm due to such misuse. For example, someone can generate a list of all gay marriages in the country and share it with a hate group. That their

DRAFT

neighbors already know one is gay is not a response because the information would reach people in neighborhoods that a gay couple might specifically avoid because they do not feel safe. Or, consider the risks of a fetishist pedophile using Census data to help generate a specific list of, say, children belonging to some particular ethnic, gender, and age group.
.

III. Accuracy

Congress, with U.S. Supreme Court support, has so far decided to take the Constitution's call for an "actual Enumeration" rather literally (https://supreme.justia.com/cases/federal/us/525/316/). The phrase "actual Enumeration" brings to mind counting the chairs in a room or the books on a shelf. (See, for example, the government's discussion of imputation as described in the Supreme Court decision on UTAH V. EVANS (01-714) 536 U.S. 452 (2002) 182 F. Supp. 2d 1165, affirmed.) But the "true" population of a state, city, or block at 11:59 pm on April 1, 2020 is not perfectly well-defined because people are born or die every minute of every day. In addition, measurement challenges arise because others change residence, immigrate, or emigrate.

Or consider non-response rates. Most policy, business, and research users of Census data do not adjust for the statistical estimates of undercounts routinely produced by the Census Bureau itself. Furthermore, the entries that appear in the Bureau's official tables are not the simple sums of data originally reported from the field and collected in the "Decennial Response File." Instead, the "Census Edited File" is the result of Census staff's de-duplications, imputations, and other forms of "pre-processing." That file, in turn, has historically been subjected to ad hoc "swapping" and other informal disclosure avoidance mechanisms in order to produce the "Hundred-percent Detail File" on which tabulations are actually based. For example, top and bottom-coding has been used to hide extreme values of certain variables. While introducing bias for many estimators and, in particular, making the population seem more homogeneous than it really is, all this tinkering may render the published totals more useful in many respects but less so in others. It also illustrates, in any case, why there could be many different and defensible estimates of the population resident in a given geographic subdivision on April 1.

What is new? New privacy protections will force broad recognition that official "counts" are estimates, not pure truth. Such numbers, regardless of how accurate they actually are, have always played an important legal and regulatory role. The Consumer Price Index (CPI) is similar in this regard. There may be better ways of measuring inflation for particular purposes, but many contracts and calculations depend critically and conveniently on the CPI exactly as announced each month by professionals at the U.S. Bureau of Labor Statistics. Everyone trusts that the procedures for producing such numbers, though highly technical, are also consistent, reasonable, transparent, and not readily subject to manipulation (Porter 1996). By way of contrast, counting up people, chairs, or books sounds like a task anyone ought to be able to do. Not only that, but the results seem like they should be the same no matter what. Because of this misperception, and because we live in a society where data, lawsuits, and special interest

DRAFT

groups have become so pervasive, the "accuracy" of the Decennial Census is now much harder to define and much easier to challenge.

Introducing formal privacy protection to Decennial Census data products may ultimately lead to changes in how laws and regulations are written. For example, knife-edge eligibility or allocation criteria for policies and programs will likely seem more arbitrary once everyone is more cognizant of how these get applied to population estimates only and not to literal counts. Federal programs where eligibility can hinge on small differences in Census products include Housing and Urban Development Community Block Grants, Rural Business Development Grants, and the Rural Micro-entrepreneurship Program. Federal funding allocations that can depend sensitively on minor differences in Decennial Census counts include the FMAP (Federal Medical Assistance Percentage), Health and Human Services Social Services Block Grants, and multiple USDA rural development programs. Of course, as many have noted, there was always noise (mostly with unknown properties) in Census tables. Now the transparent injection of noise with known properties may provide an impetus to design policies less sensitive to the inaccuracy in all statistical indicators, whatever the cause. Phased benefit levels and eligibility may become the rule, as opposed to sharp criteria that flip on or off based on a "count" difference that is statistically meaningless.

More generally, there is growing awareness that the Decennial Census is not necessarily a literal "census" as that term is used in statistics textbooks. Rather than the result of a simple and complete enumeration of the population, it is based instead on a particular and idiosyncratic sampling of the actual population. Every effort is made to conduct a comprehensive count, of course, but it is not. Neither is the Decennial Census a random sample either. Inevitable statistical biases, owing for example to systematic differences in the counting of certain groups, are routinely studied by the Census Bureau and should be of more interest to researchers or decision makers who have often found it more convenient to take the published tables at face value. Basing conclusions on the properties of a particular sample rather than on properties of the population from which it is drawn constitutes the kind of major error that, arguably, much of modern statistical theory has been specifically designed to avoid. Some applications of the Decennial Census findings are, from this point of view, egregious cases of overfitting to the sample that therefore have little if any statistical significance or justification. (See Groves and Lyberg 2010.)

IV. Privacy

The Decennial Census asks all residents for basic demographic information (age, sex, race, and Hispanic or Latino origin) as well as housing information (household size, composition, and occupancy tenure as owner or renter). To reassure respondents about their privacy, the Census Bureau cites Title 13 of the U.S. Code. Section 9 requires that neither the Bureau in particular, nor the Department of Commerce generally, shall:

T1. Use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or

T2. Make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or

T3. Permit anyone other than the sworn officers and employees of the Department, bureau, or agency thereof to examine the individual reports (see https://www.census.gov/about/policies/privacy/data_stewardship/title_13_-_protection_of_confidential_information.html).

Researchers who need direct access to sensitive data must first take on "sworn status." That is, they must pledge to obey Title 13 for life, or face the same significant penalties applicable to government employees. They may then be granted permission to work "behind the firewall" in a highly controlled Federal Statistical Research Data Center (FSRDC). But for most researchers, not to mention more casual users of Census results, both aggregate statistical tables as well as public-use microdata samples are easily available and suffice for many purposes. These releases are intended to comply with Title 13 by being subjected to established "disclosure limitation methods" governed by the Census Disclosure Review Board. Traditionally, this is a loosely documented process that, according to a handbook for statistical agencies (Data et al. 2001), consists of three steps:

H1. Information that directly reveals the identity of the respondents is suppressed.

H2. Information that may indirectly reveal the identity of a respondent is suppressed. This can be accomplished by reducing the variation within the data through rounding, top- and bottom- coding, collapsing response categories, and suppressing information such as detailed geography.

H3. Some uncertainty can be introduced into the reported data. This can be accomplished by altering the underlying data through swapping of reported values among similar respondents, *adding predetermined random noise to the data, and performing other more structured randomization of the data* [emphasis added].

What is new? Due to advances in both theoretical understanding and practical computing power, the informal disclosure limitation methods listed above no longer provide adequate privacy protection. Beginning in 2003, a series of mathematical results established what is often called the Fundamental Law of Information Recovery (Dinur and Nissim 2003; Dwork and Roth 2014). Imagine a curator who keeps a database behind a secure firewall, and answers statistical questions about it submitted by an analyst. If the curator answers too many of those questions with too much accuracy, the theorem shows how an analyst can, with high probability, exactly reconstruct every bit of the database. Researchers who worked with Census data in the past have known that confidential information about individuals could sometimes be reidentified. But it was assumed that the scope and significance of this risk was limited enough to be overcome by swapping, cell suppression, aggregation, top-coding, and other informal methods. Recent tests have proven otherwise. Starting only with the 2010 tabulations

DRAFT

as made available to the public, here is what "white hat" Census researchers were able to discover about individuals (Abowd 2018b):

A1. All five of the variables studied—block, sex, age in years, race, and ethnicity—were reconstructed correctly for 46% of the population (about 142 million people). The Census block variable was correct for everyone. Since birthday records relative to April 1 are not always easily resolved, allowing age to vary by plus or minus one year brings the proportion of correctly reconstructed records up to 71% (about 219 million people).

A2. Combined with data commercially available in 2010, these reconstructed records could be uniquely matched to names, addresses, and many other fields of personal information for 45% of the total population. Block and sex matches were exact, while age was allowed to vary by plus or minus one year.

A3. Finally, 38% of these putative matches (about 52 million people) could be verified as accurate re-identifications by checking the name, block, sex, age, race, and ethnicity against the confidential Census data. (Note that, without access to that confidential data, it may not be easy to tell which of the putative matches are actually correct.)

No one in 2010 would have imagined undertaking such an exercise without the theorems and computational capacities that have only recently become available. But once the possibilities are understood, it is not so hard. For example, Mark Hansen's journalism graduate students at Columbia subsequently studied Manhattan rather than the whole country. (See https://www.nytimes.com/2018/12/05/upshot/to-reduce-privacy-risks-the-census-plans-to-report-less-accurate-data.html.) Using standard software and Census tables, they were able to start generating similar reidentification results in less than a week. Note that the Census Bureau releases billions of numbers in its tables, but there are less than 350 million people in the U.S. So, by using some linear algebra, it is not that hard to determine when there is one and only one person with a certain set of identifying characteristics like age, gender, location, etc.

One objection to the Census approach, even from those who would agree about the importance of counting everyone, argues that the potential damage from reidentification is minimal because the information the Census collects is not very sensitive. Few making such arguments, however, belong to vulnerable populations such as those at risk of being found by an abusive partner, hate group, being deported by the Immigration and Customs Enforcement Agency, or being evicted by housing authorities. Some hold that reconstruction of the data is not the same as identification, and that the Census Bureau should not be too concerned if its data can reveal individual's attributes (Ruggles 2019). But this assumes incorrectly that we know all the ways the information could be used. Hackers, not to mention commercial vendors, can easily process or link with Census data to the detriment of respondents. In the future, perhaps their actions should be declared illegal and prosecuted. For now, however,

DRAFT

what matters is the clarity of current legislation that prohibits Census from providing the public with information that identifies respondents.

For example, the methods that recovered race and ethnicity information for millions of residents could also be used to reveal citizenship information. One way would be for the Census Bureau to include such counts as part of their CVAP (Citizens Voting Age Population) file. Alternatively, outsiders could also match records using artificial intelligence techniques. (For information about the CVAP, see https://www.census.gov/programs-surveys/decennial-census/about/voting-rights/cvap.html.) Given these stakes, there are always incentives for some groups to discourage participation in the Census by other groups. McGeeney et al. (2019, page 39) shows that communities of color are concerned about reidentification, introducing the risk of a chilling effect on participation for those who see themselves as data vulnerable.

To be clear, there have been no recent examples of researchers bringing harm to an individual based on personally identifiable information recovered based on Census data. This is a point of pride for everyone involved. For it to remain so as circumstances change requires vigilance, however. One incident of data misuse—or even rumors of potential misuse—could easily jeopardize the government's ability to carry out its Constitutionally mandated counting duties ever again.

Federal officials must, in any case, interpret and implement laws like Title 13 impartially and with integrity. Others who, in contrast, do not face fines, jail, profiling, disenfranchisement, or deportation if they are wrong may have the luxury of discussing at length what the word "identified" could and should mean in that statute. The only alternative offered by critics so far has been to go back to opaque, discredited, and *ad hoc* methods of privacy protection such as swapping. Disputes like this about how to apply laws in light of changing societal expectations and technological capacities are typically settled in court. Until then, researchers and other data-users simply need to understand the systematic privacy protections that professional Census Bureau officials are instituting.

V. Privacy vs. Accuracy

The re-identification attacks just described suggest that the Census Bureau would violate Section 9 of Title 13 if the agency calculated and released tables for the 2020 Census in the same way it did for the 2010 Census. Such a conclusion does not depend, for example, on whether data about an individual's immigration status is obtained by posing the question on Census forms or whether it is determined from administrative records instead. (For information about complying with the Executive Order on this, see doi:10.1126/science.aaz5220.) To continue relying either on forbearance or on cell suppression, swapping, and the like to protect millions of people's sensitive information would be irresponsible and impractical to carry out. Moreover, such informal methods provide neither privacy guarantees nor measurements of how much accuracy is sacrificed.

DRAFT

What is new? The first and, so far, only approaches to *formal* disclosure limitation are all based on an idea introduced in Dwork et al. (2006). Early papers in this field explicitly treat the release of Census tables as a motivating problem (Barak et al. 2007). Consistent with the Information Recovery Theorem mentioned above, even aggregate statistics in such tables diminish both privacy and validity protections. The conceptual framework for specifying, measuring, and controlling such leakage is called "Differential Privacy."

Despite the name, Differential Privacy is not about providing different levels of privacy protection to different people. Nor does the term refer to any one particular algorithm, but rather to a property that some disclosure avoidance algorithms possess. Note also that being protective of privacy is not, on this account, a property of the dataset but rather a feature of certain methods used to release statistics calculated from a dataset.

The theory has two parts. One is a definition of what it means for a "query mechanism" to satisfy "ε-differential privacy" where, as usual, the Greek letter epsilon stands for a small positive number. The other is to establish key properties implied by this definition and illustrate them by producing practical examples of algorithms that meet the stated criteria. For a primer that introduces the most important technical details about Differential Privacy, see the Appendix to this paper. Refinements and variations are noted there as well, including some being implemented by the Census Bureau, but we stick with the basic notions for now.

To begin appreciating the implications for researchers and data scientists, imagine a sensitive dataset kept by the curator mentioned before, who works behind a secure firewall that functions as a privacy barrier. Analysts pose questions, but the curator only returns to them results produced by running a "query mechanism" on the data. For example, the analyst may ask for the average age of women living in small geographic area. This query mechanism will produce a result based on the confidential data but treated in some way to be consistent with preserving a certain degree of privacy determined by ε. Ideally, the curator would like to ensure that the analyst cannot even find out whether any given person is in the dataset, let alone anything else about that individual. That is because, if I am deciding whether to allow use of my information, I may care not only about what an analyst could find out about my characteristics but also about whether anyone could determine whether I am listed at all. In other words, the curator should limit how well an analyst can distinguish between "neighboring" datasets, i.e., pairs that are identical except that there is a single person listed in one but not the other. (The same approach also works using other criteria for when datasets should be considered neighbors, such as differing only in the *attributes* of one person. See Kifer and Machanavajjhala [2011] for more details about bounded vs. unbounded differential privacy.)

Roughly, a query mechanism is said to satisfy "ε-differential privacy" if the answer it gives cannot change the analyst's prior odds about whether or not any given individual is in the dataset by a factor that differs from 1 by more than ε. So the smaller the ε, the less you can learn about individuals from a query mechanism that satisfies ε-differential privacy. For a data generating process like census-taking, you can think of ε as measuring "plausible deniability."

DRAFT

That is, if I want to claim that my data is not even in the dataset at all, ε indicates the likelihood that an analyst could verify or falsify my claim based on a query made with a mechanism satisfying ε-differential privacy. (This does not, of course, prevent my being harmed by statistical findings calculated from a given dataset. But as long as the calculations are carried out by such a query mechanism, those findings would hardly have been any different regardless of whether my information was in the dataset or not. So precisely what it does protect against is *participation* risk. And regardless of how you think officials should weigh various re-identification threats *ex post*, this is precisely the *ex ante* privacy risk to minimize in order to maximize participation in the Census. Arguing over the meaning of the reconstruction attack would hardly have reassured residents who were wondering whether to fill in Census forms out of fear that the government or other actors might use the personal information asked for to track, deport, evict, impugn, or discriminate against them.)

Do such query mechanisms actually exist? Yes, they do. It turns out that you cannot, for example, allow the analyst to receive precise answers to statistical questions—even innocent looking ones like averages or percentiles. If the analyst wants to know such a statistic, the curator calculates it behind the firewall, but then must add a small amount of random noise to the answer before returning it to the analyst. The parameter ε governs the inevitable trade-off between privacy and accuracy. More privacy protection goes along with smaller ε, but adding more noise also provides less accuracy. In the limit as ε grows infinite, the noise recedes to reveal the unprotected statistic as calculated from the confidential data.

What do we mean by ‘‘adding more noise to" or, more colloquially, "fuzzing" an answer to a statistical question? We mean adding a new term to the unprotected answer. That term is a random variable. Its value can be drawn from a distribution whose density function and parameters are announced publicly. The Laplace distribution is a convenient choice, for example, and the corresponding query mechanism is called the Laplace mechanism. That distribution will be highly concentrated around zero (to minimize distortions), have mean zero (to avoid introducing bias), and have variance inversely proportional to ε^2 (to properly protect privacy). So you can think of more noise as just a way of saying more variance.

Note that this discussion so far is specifically about fuzzing *one particular statistic* calculated from that data before release to the analyst, rather than the alternative of blanketing the *entries* in the original dataset with random perturbations. (The latter approach is also possible, but is distinct from the method adopted for the 2020 Census and goes under the name "local differential privacy.") Still, researchers should have at least two immediate and serious concerns about this approach.

First, random noise that is supposed to be small and unbiased can nevertheless take on large values sometimes. What if, for example, the mechanism answering a population count query returns a negative number? Because the workings of the mechanism are entirely transparent, we will argue that researchers who are always dealing with noisy data anyway can cope with this, too. The public, however, quite reasonably expects Census tables to have nonnegative entries. If a noisy measurement needs further adjustment to ensure face validity like this, what

DRAFT

happens to all those privacy guarantees? When strictly applied, the Post-Processing Theorem has this covered (see Appendix). It says that if, after first running a query mechanism that satisfies ε-differential privacy, a curator performs other random or deterministic operations on the answer before releasing it, the combined query mechanism will still satisfy ε-differential privacy *as long as the later operations do not refer back to the original dataset again in any further way*. This means that, in contrast with other disclosure avoidance techniques, the formal guarantees provided by differential privacy do not erode no matter what new datasets or computing capabilities become available in the future.

A second concern is that researchers usually ask about more than one statistic. The Census Bureau has to calculate billions, after all, to fill the tables it publishes. Note that, in order to protect privacy as described above, each one of those table entries must be determined by running a query mechanism satisfying ε-differential privacy for some value of ε or another. But surely releasing more and more information about a dataset makes it harder to protect the privacy of individuals listed there? It does, but the theory of Differential Privacy allows for precise accounting and control of how its formal guarantees are affected as analysts ask more and more questions. This is again in marked contrast with other, more informal, approaches to disclosure avoidance.

Specifically, the Composition Theorem says that, if a dataset curator runs a query mechanism once that satisfies ε_1-differential privacy and runs another query mechanism once that satisfies ε_2-differential privacy, then the combined mechanism that releases both results satisfies $(\varepsilon_1 + \varepsilon_2)$-differential privacy. So a curator committed to offering privacy guarantees corresponding to some ε^* but who also wants to handle n queries must make sure that $\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_n \leq \varepsilon^*$. In other words, such a curator has to manage a "privacy-loss budget" that gradually gets used up by responding to each new question posed by the data analysts.

Importantly, the appropriate value of ε for a given situation is not a mathematical question, but rather *a policy decision* for the curator to make and maintain. The Bureau's Data Stewardship Executive Policy Committee is the body responsible for setting the overall value of ε^* that governs release of 2020 Census information. Their decision will have notable impact on the research community. For example, a value of 0.01 implies that, if an analyst had even odds about whether a given individual is in the dataset or not, her posterior odds after receiving an answer to a query could go from 50:50 to about 52.5:47.5. (In other words, the posterior probability could change by about 2.5 percentage points.) The value used by the Census Bureau in its 2018 End-to-End Test in Rhode Island was $\varepsilon = 0.25$. Some commercial software defaults to an epsilon setting of 1, which means that a query result could transform even odds into odds of almost 3 to 1. Setting epsilon equal to 3 is another common choice among commercial providers. Especially if an analyst's prior odds about whether I am included in a given data set are one in hundreds of millions, then even an epsilon above 10 can provide some meaningful measure of privacy protection. Academics have sometimes argued that simply having a finite epsilon is more important than its precise value since such a policy prevents a query, or series of queries, from revealing with certainty whether an individual is or is not present within the dataset under study.

DRAFT

On June 9, 2021, the Census Bureau's Data Stewardship Executive Policy Committee announced a total privacy-loss budget for the redistricting data product and its subsidiary population and housing products. The overall value of ε^* for 2020 Census information is now set at 19.61, which includes ε=17.14 for the persons file and ε=2.47 for the housing unit data. For full information, see https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html. Note also that the interpretation of this value also depends on the implementation of Differential Privacy. Relaxed variants, including concentrated Differential Privacy, are discussed briefly in the Appendix below.

VI. Histograms, Statistical Priorities, and Strategy Matrices

We normally think of data usage as having zero marginal cost. Differential Privacy, as a conceptual framework, teaches otherwise. In fact, producing and protecting useful Census tables turns out to be much more like any other microeconomic problem. There are, as we have seen, budgets, tradeoffs, and other constraints to take into account. As when firms make their decisions, the tradeoffs are not always easy to determine in detail. We may know how to measure one variable, in our case the privacy guarantee corresponding to ε, but still wonder about how increasing or decreasing it affects traditional measures of statistical accuracy on the margin. As we will see, there are active mathematical explorations underway concerning the "technology frontier" that delineates which combinations of privacy and accuracy are feasible. As an engineering matter, though, one practical rule is familiar to those who routinely work on confidential or secret matters: only deal with data you need.

For what does the Census Bureau need decennial data? A primary responsibility is to produce lots of tables, of course. In 2010, for example, the Summary File 1 data product (called the SF1) contained hundreds of tables for those who want to know how many US residents there are with various characteristics. The 2020 products will differ in various respects from the 2010 products. (See the Census Bureau's "crosswalk" files for planned correspondence between 2010 and 2020 products. https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/2020-census-data-products-planning-crosswalk.xlsx.)

Tables based on the Decennial Census traditionally include counts and cross-tabulations regarding age, sex, race, Hispanic or Latino origin, household relationship, household type, household size, group quarters population, whether a housing unit is occupied or vacant, and whether the occupied housing unit is owned or rented. Data are available down to the Census block level in many, but not all, cases. Note that, especially since millions of those blocks have less than ten residents, household relationship may be a sensitive matter for LGBT families, or household size could be a legal issue for those living in overcrowded conditions, for example. Note also that the Decennial Census does *not* ask about ancestry, educational attainment, income, language proficiency, migration, disability, employment, or housing characteristics.

DRAFT

Those questions do appear on the American Community Survey that replaced the so-called "long form."

The Census Redistricting Summary File is often referred to as the "P.L.94-171," after the Public Law of that number which mandates the release of tables broken down by race, by Hispanic or Latino origin, and by age (under or over 18) as needed by the states to draw legal voting districts. Though only required to provide population counts, traditionally the Bureau has agreed to tabulate these other variables down to the Census Block level. All told, the P.L.94-171 tables contain nearly 2.8 billion entries. In 2010, the Summary File 1 (SF1) tables included nearly 2.8 billion more statistics in addition to those. Its successor, the Demographic and Housing Characteristics (DHC) File, will as well. In all, the number of table entries that the Bureau releases is estimated to exceed 150 billion.

As usual, the first steps in preparing these products will include processing to remove duplicates, to impute values that replace non-responses, etc. The result is the Census Edited File (CEF) that must then undergo disclosure avoidance procedures to create the file from which tabulations are calculated. In the past, that file was internally referred to as the "Hundred Percent Detail File." It was the basis for creating tables that were deemed releasable after disclosure avoidance techniques such as "swapping" were applied. In 2020, a much more rigorous and computationally intensive Disclosure Avoidance System will create the input file for data product creation. Called the "Microdata Detail File," its design reflects the advantages and disadvantages of producing data that can feed into existing systems for generating tables more or less as they were used in 2010.

What precisely will be new? As discussed above, the Census Bureau has determined that publishing the P.L. 94-171 and other tables without formal privacy protections would violate its legal mandate against releasing data that can be tied back to individuals. Table entries will therefore be computed using query mechanisms that satisfy ε-differential privacy. The application of such techniques is greatly facilitated by knowing in advance which statistical questions need to be answered publicly—even if, or especially if, there are billions of these calculations to make, as is the case when preparing Census tables.

The value of ε^* fixed for the release of 2020 Decennial data as a whole determines a privacy-loss budget that, as discussed above, will be used up over time as the Census releases responses to more differentially private queries. One strategy the Census Bureau will not use is to reach into the Census Edited File (CEF) and randomly change the ages, zip codes, or other information recorded there. To be clear, that is neither required nor recommended to satisfy Differential Privacy. Nor is it even remotely practical. Just imagine replacing the unprotected value of the i^{th} data entry with the value returned by a query mechanism satisfying ε_i-differential privacy. Each such change would subtract ε_i from the privacy-loss budget. Because the index i runs into the billions, either the privacy-loss budget would have to be huge—corresponding to very little privacy protection—or many of the ε_i would have to be tiny—corresponding to very little accuracy.

DRAFT

Another strategy the Bureau will not use is to start by applying privacy protection to Census block data, then simply aggregate up from there to produce tables for larger geographic regions like counties and states. The reason is that this would also introduce too much noise. Of course, noise tends to cancel when averaged. But such intuition does not apply here. Rather than averages, most Census table entries are sums. And by the Variance Addition Law, if you add up several independent random variables, the variance of the sum is the sum of the variances.

Instead, the Census Bureau's strategy is to adjust histograms as follows. For simplicity, we can start by considering only the traditional P.L.94-171 tables in this section that describe the nation as a whole, ignoring geographical or other distinctions at the state and lower levels. There are 1,227,744 possible combinations of the remaining variables: 2 for sex; 2 for Hispanic or not; 63 for race; 42 for relationship to householder; and 115 for ages from 0 to 114. For each of these bins, count how many people fit the corresponding description. The resulting histogram is a convenient way of representing the national microdata. From it, you can make all the national tables you want, including aggregate statistics like those cross-tabulations in P.L.94-171 files about how many U.S. residents of some sort also have some other characteristic. Such a histogram is therefore said to be "fully saturated."

Creating and releasing all these Census tables is, of course, a "dissemination-based" access strategy. This is often contrasted with "query-based" access in literature about confidentiality (see Kinney et al. 2009), meaning that users interact directly with the data. In our discussion of "query mechanisms," however, we have instead been imagining how the Census Bureau itself can produce a single privacy-protected statistic that might then, for example, become an entry in one of the tables it disseminates. Doing this once or twice is not so hard, as we have seen. Now the problem is how to do this for the billions of table entries the Bureau needs to produce without letting noise overwhelm all the measurements. Working with a fully-saturated histogram, as just described, is a convenient place not only to begin organizing the calculations but also to begin seeing how various choices and priorities affect the privacy-loss budget.

So how can the national histogram, or any such table for that matter, be safely and usefully brought over the privacy barrier for public use? The Bureau could, for example, protect privacy by adding some random noise to each of the bin counts, requiring 1,227,744 queries to be executed (one for each bin) that each satisfy ε-differential privacy. That sounds like a big hit to the privacy-loss budget if we have to sum up the epsilons associated with fuzzing each bin count. Remarkably, we can do much better than that. The key is that each U.S. resident counts in at most one bin. Going back to the definition of ε, that means that an analyst's prior odds about whether a given person even participated in the Census at all could change, after the whole exercise is done, by a factor that is no higher than the factor by which the analyst's odds would change about whether the individual is in any one of the bins. This follows in general from the "Parallel Composition Theorem" as described in the Appendix. (See also there the definition of "sensitivity" and verify that it not only determines the constant of inverse proportionality between the noise variance for a query and ϵ^2, but also that it conveniently equals one for counting queries about histogram bins.)

DRAFT

So if the query mechanism for each bin is ε_0-differential privacy for some fixed $0 < \varepsilon_0 < \varepsilon^*$, for example, then preparing a privacy-protected histogram this way would use up at most ε_0 of the total privacy-loss budget ε^*. In fact, if different epsilons are used to fuzz each bin, then the overall ε^* will be the maximum over all of these.

So far, so good. But this strategy still involves lots of noise. Adding up the bin counts to calculate the total U.S. population, for example, would give the correct answer plus a random variable whose variance would be over a million times the variance of the random variables added to each bin. Can we do better?

One guiding principle is to avoid applying privacy protection directly to any variable not strictly needed to produce the tables that will be released. Finer granularity helps with neither privacy nor accuracy. Keeping that in mind, notice that some of those table entries can be computed by taking linear combinations of others. If interested in the one particular category in the population that breaks down by gender, for example, there is no reason to add noise to the number of males in that category and to the number of females, and then to the total number of people in that category as well. In fact, that method would produce fuzzed male and female numbers that no longer sum to the fuzzed total. Protecting any two of the three variables will suffice instead. When you do compute the third from the other two, however, note that it will have twice as much noise as they do (again by the Variance Addition Law). So, it pays to be careful about which variables you inject noise into first and which you derive from them. High priority statistics and tables should be protected first since those measurements will be less noisy.

Good idea, but implementation raises two challenges. One is how to find out which statistics or tables really are more important to users than others. Another problem, given that information about priorities, is organizing such calculations when dealing with millions of variables and billions of table entries instead of a handful as in the gender example above. There we eventually wanted all three fuzzed numbers corresponding to a male count, a female count, and the total count. This is our "workload." The "strategy" refers to our decision about which two variables to prioritize by fuzzing them first so that the third fuzzed number can then be calculated by addition or subtraction. Producing even more complicated table entries also involves linear operations only. So the full-blown task can be organized and carried out in a way that satisfies ε-differential privacy using what is called the Matrix Mechanism (see Li et al. 2015).

Briefly, the Matrix Mechanism works this way. Order all the bins and imagine writing down the unprotected number from each bin to form a long column vector x. A count needed for a table can be represented as a row vector w whose components are either 1 for a bin that contributes to the total or 0 for bins that do not. Now list all those vectors as rows in a huge matrix W. It is called the "workload matrix" because Wx is a vector listing all the unprotected table entries

DRAFT

ever needed. You do not want to add noise to each of them (way too many) and you would prefer not to add noise to each entry of x either (still too many). Instead, imagine factoring the workload as $W = UA$ where A is a smaller and carefully chosen matrix. It is called the "strategy matrix" because the shorter list of components in the vector $y = Ax$ are supposed to be both high priority table entries individually and, as a group, all you need to eventually compute every other table entry. Now just add noise to y and apply U (which, for example, can be taken to be WA^+ where A^+ stands for the Moore-Penrose inverse for A). The result will be a list of all table entries desired, but now privacy-protected, using less noise overall than other methods and, in particular, even less noise for the high priority components specified when selecting the strategy A.

To solicit priorities from users, the Census Bureau monitored which tables seem to be used more frequently than others. The Bureau also issued a request in the *Federal Register* asking the public about critical use-cases for Decennial data (https://www.federalregister.gov/documents/2018/07/19/2018-15458/soliciting-feedback-from-users-on-2020-census-data-products.) Neither the responses from data users generally nor those from the research community in particular were initially as informative as was hoped for planning purposes. This was, after all, a new method of soliciting input from Decennial data users. Many addressed the utility of data from the American Community Survey instead, perhaps because there have been calls from members of Congress and others to eliminate altogether this "long form" sampling of the U.S. population. A few voiced concerns about setting ε^*. But another kind of input the Census Bureau especially needs is advice about determining strategy matrices. That includes the A we have discussed to improve national tables as well as analogues discussed below for state, county, and lower level tabulations.

More interaction with data users about table priorities for the Decennial results would undoubtedly be beneficial to all. The Bureau is actively engaging in these discussions and refining its disclosure limitation procedures based on user feedback. Of course, statistics that are of great importance to some users are of little importance to others. For a fixed privacy-loss budget, however, more accuracy in one domain means less elsewhere. So once again, Census officials must contend with inevitable trade-offs. Some relatively inessential tables may therefore be eliminated altogether. Regardless of how assiduously the Census Bureau tries to address input from users about their priorities, *any* method of protecting privacy—whether based on Differential Privacy or not—is bound to produce results that are acceptable for certain use cases but disappointing for others. Especially when studying small groups, researchers should be aware that counts of people with unusual characteristics may be altered in ways that, though alarming for particular purposes, are nevertheless necessary to protect privacy. The theory of Differential Privacy at least gives a conceptual framework for making such decisions strategically, explicitly, and accountably.

VII. The Noisy Measurement File, Face Validity, and Post-Processing

DRAFT

Once the specific strategy for injecting noise is set, the Census Bureau can produce what it calls the "Noisy Measurement File (NMF)." It has great statistical value because the tabulations there will be unbiased by the addition of mean-zero noise whose distribution is entirely known. Researchers who work with this file can take the properties of the added noise into account in their analytical approach.

Many users of the data are not researchers, however, and have other needs. The public has certain expectations about what a published Census table should look like. In particular, they expect table entries to conform to the logical constraints found in the protected enumerated data. Counts of people, for example, should not be fractional or negative. Fuzzed data do not necessarily have that kind of face validity. Indeed, unrealistic features in that file could render it inappropriate for some purposes and confusing enough to sow distrust in (and even inspire ridicule of) the Decennial Census among the general public.

To eliminate unsettling entries from its official publications, the Census Bureau will "post-process" the Noisy Measurement File. As in the previous section, we can start for the sake of simplicity by considering only the national level tables needed for the PF74-191. These post-processing calculations can be arranged to avoid accessing the original private data again, and so do not draw further on the privacy-loss budget. We might sometimes refer to the changes made after noise infusion as "cosmetic," to be evocative, not pejorative, and to emphasize that the Noisy Measurement File has better statistical properties and research uses than the tables made to look like they report simple counts.

Post-processing addresses at least six serious challenges to the face validity of NMF tabulations.

C1. **Negative Counts:** Adding noise to a count can make it go negative. This is not a good look. How do you explain negative people to the general public? The planned remedy is to apply a Non-Negative Least Squares (NNLS) algorithm that, for a given table with negative entries, finds another table with non-negative entries that is closest to the original as measured by the sum of the squares of the entry differences. Like all post-processing, such a projection procedure does not draw on the privacy-loss budget. But it does, of course, introduce a positive bias in many statistics that will be most apparent among counts that were small to begin with.

C2. **Structural Zeros:** Two-year-olds and other small children are not supposed to head households. Any count that, after noise injection, suggests otherwise should ideally be overridden. Examples like this can be built into the NNLS algorithms as additional constraints along with the non-negativity requirements. Each adds to the considerable computational burden, however.

C3. **Fractional Counts:** Again, not a good look. Users expect Census tables to report whole numbers of people. Laplacian noise, by contrast, can take on any value. By instead employing the closely-related Geometric Mechanism, formal privacy protection can be achieved without introducing any non-integers. Nevertheless, fractional entries sneak back in at least two ways, via the Matrix Method and NNLS algorithms. So, a second stage of post-processing is required to round everything back to whole numbers.

DRAFT

C4. **Invariants:** The Census Bureau has announced that a few of the important numbers computed behind its firewall will be taken over the privacy barrier and released to the public directly without the injection of noise. These are the Bureau's best total population estimates for both the nation as a whole as well as each of the states and territories, the number of housing units at the block level, and the number and type of group quarters facilities at the block level. (See https://content.govdelivery.com/accounts/USCENSUS/bulletins/2ae5eda.) In principle, this voids the formal privacy guarantees provided by Differential Privacy. But the policy applies only when faith in these numbers is especially critical (as in reapportionment) and when the privacy implications are deemed minimal (as in large counts such as state populations). The other invariants (the number of units or group quarters within a Census block, although not their populations) are included on the grounds that such counts are, in principle, observable by the public. Concerning privacy guarantees in the presence of publicly known statistics, see also Ligett et al. (2020), Kifer and Machanavajjhala (2011), Abowd et al. (2019), and Gong and Meng (2020).

C5. **Logical Consistency:** More generally, the count of some subcategory should not exceed the count in its parent category, and the sum over an exclusive and exhaustive set of subcategories should equal that of the parent. When it comes to basic geographic hierarchies, the "TopDown Algorithm" described below will ensure that, for example, the county populations within a state add up to that of the state. But there are other constraints whose violation might strain credibility, such as a household that reports more residents than the block that contains it. Again, noise injection need not respect such constraints unless post-processing re-imposes them.

C6. **Limiting Behavior:** As noted earlier, the pre-post-processed entries in the Noisy Measurement File will gracefully converge to their unprotected predecessors as the parameter epsilon tends to infinity and the injected noise subsides. Ideally, whatever post-processing system the Bureau implements should also produce tables with the same asymptotic property regarding convergence to their unprotected values.

Accomplishing these six goals places extraordinary demands on the post-processing system. There are massive quadratic programs to solve, for example, with both integer as well as many other constraints. Each computational run requires significant time, expense, and storage using state-of-the-art commercial software, currently supplied by "The Gurobi Optimizer" (https://www.gurobi.com/products/gurobi-optimizer/). The process is opaque, proprietary, irreproducible, and—even with all that technological firepower—not guaranteed to produce an optimal solution. When it fails, the protocol is to begin relaxing some of the constraints described above until tables are generated that look suitable for public release.

Experience so far with these techniques has led the Census Bureau to several conclusions. One is that "Post-processing error tends to be much larger than differential privacy error." Another is that "Improving post-processing is not constrained by differential privacy" (see https://www2.census.gov/about/policies/2020-03-05-differential-privacy.pdf). Intense work is

therefore ongoing to refine and revamp the procedures for making tables presentable enough to be trusted by the public.

For researchers, though, there are strong arguments to avoid the uncertain impact of post-processing on the validity of statistical inferences. Professionals should have little trouble interpreting negative or fractional counts as remnants of the fully transparent process for protecting privacy, and would much rather do that than cope with unexplainable distortions introduced only after all the privacy protection is complete. The Census Bureau has yet to decide how or if the Noisy Measurement File will be available.

VIII. Geographic Hierarchy, Protected Microdata, and Reconstruction Revisited

Having concentrated on the construction of national tables in the previous sections, the question remains about how to consistently construct privacy-protected tables about finer geographic divisions. This will actually proceed in stages, from the national level to the states, from the state to the county, etc. Post-processing will take place at each stage rather than all at once. The entire procedure, known as the TopDown Algorithm, can be viewed as making surprising use of the microdata reconstruction techniques discussed above.

Recall that, starting with tables that look like they could appear in the P.L. 94-171, the Fundamental Law of Information Reconstructions shows how to recover the microdata about individuals necessary to generate those tables (see Dinur and Nissim 2003). Why not carry out the very same steps, but start with data products that have been privacy protected and post-processed for release? The Privacy-Protected Microdata Files (PPMF) backed out from such tables will look like complete records of individuals. And when aggregated, the characteristics of these imaginary individuals will exactly reproduce the released tables. But, by the Post-Processing Theorem (see the Appendix), those records can never reveal anything more about real individuals than the fuzzed tables from which they were derived. Depending on the epsilon values used when preparing those tables, of course, their entries can be made relatively close to the unfuzzed counts.

Keeping in mind this method of manufacturing microdata, let's go back to the national tables. We previously saw how the Census Bureau will construct and protect these. Suppose that bringing national level tables over the privacy barrier uses up ε_1 of the total privacy-loss budget ε^*. Constructing privacy-protected and fully releasable microdata corresponding to that information entails no further budget drain. The results will comprise records for about 330,000,000 imaginary people, but without any names, addresses, or other geographic data.

That is an interesting construction at the national level, but how can it help at the state level? Setting aside temporarily the invariance of state population totals, pick a value of ε_2 that each state (or state equivalent) can use to produce privacy-protected tables and auxiliary information about its population. By the Parallel Composition Theorem, this will only use up ε_2 of the total privacy-loss budget ε^*even after all the states are done. Thus, all that the Census

Bureau needs to do is distribute the 330,000,000 records among the different states so as to both: a) match the fuzzed state tables as closely as possible and b) satisfy the six face-validity criteria discussed above as closely as possible. This closeness can be expressed either in terms of average absolute error or mean squared error. Combinatorically, there are zillions of possible ways of associating each of those records with a state. But with clever algorithms and enough patience, a good fit can be determined. As mentioned above in our discussion of post-processing, the calculations are massive enough that the Census Bureau uses a commercial system, the Gurobi Optimizer, to complete them in a timely manner.

The result will be a set of privacy-protected microdata representing the entire population, whose records now have state information that closely reproduces the state tables. Interestingly, this procedure could be viewed as a kind of swapping (on steroids) in the sense that records can be assigned to any geographic region as long as the resulting tables come out about right. Not only are the state tabulations performed on this microdata safe to release in the DHC and P.L. 94-171 files, but the microdata can also be made public, too, without any further draw on the privacy-loss budget. And, of course, state tables made this way will add up properly to the national tables, because they are all consistent with underlying microdata.

The same basic procedure carries through for standard geographical subdivisions at each lower level as well. That is, key tables and information at the county level pass over the privacy barrier using up ε_3 of the total privacy-loss budget. In some cases, the Census Bureau will also report results for geographic regions other than counties, tracts, or blocks. Intermediate calculations like this may be introduced to improve the speed and accuracy of tabulations for the traditional regions, for example. The fuzzed tables will be adjusted to improve their face validity by allocating synthetic microdata records for the state to the various counties within that state. The resulting set of microdata now has county information, it closely reproduces the original county records, it is privacy protected, and it can be used to produce protected Census Tract tables in a process that requires ε_4 of the privacy-loss budget. Similarly for Block Groups and ε_5, then finally for Census Block tables and ε_6.

Called the "TopDown Algorithm," this procedure stands in contrast with how traditional disclosure avoidance starts instead by swapping records and suppressing information at the lowest levels, then aggregating upwards. Working from the bottom up is less desirable when adding noise because, as we have seen, the variance of independent random variables adds when you sum them, and so counts would be obfuscated to a totally unacceptable degree by the time you got to estimating state or national statistics. Note that the TopDown Algorithm does guarantee that the tables from each level sum up properly and consistently to those at the next highest level. In fact, the P.L. 94-171 releases will match the usual tabulations as performed on the final set of privacy-protected microdata that includes block level information.

Like nearly everything else about the Disclosure Avoidance System for the 2020 Decennial Census, details concerning the TopDown Algorithm are still being checked, refined, renamed, and adjusted. Here are three matters of potential interest to researchers.

DRAFT

M1. The privacy-protected microdata is engineered to reproduce the counts that appear in privacy-protected tables. That does not mean this dataset will also respond well to other kinds of queries, such as nonlinear ones. Correlations, for example, could be swamped by noise and differ substantially from those computed on the unprotected microdata. Assuming that the Census Bureau makes the privacy-protected microdata available to researchers, it would be helpful to also set up validation or verification servers. By using up a bit more of the privacy-loss budget to consult the original records, such servers can check whether or not results suggested by analyzing the protected microdata are spurious.

M2. In addition to the standard geographic subdivisions, there are also "off-spine" ones such as tribal regions, congressional voting districts, or zip code areas whose boundaries do not necessarily line up with those of states, counties, or Census blocks. Many communities found preliminary test results based on the TopDown Algorithm unsatisfactory for off-spine geographies. Group Quarters such as prisons, dormitories, or military barracks also pose similar challenges. The Census Bureau is therefore developing multi-pass approaches as well as other techniques to post-process the relevant data more accurately.

M3. Low incidence measurements pose a problem more generally. Some counties have much more population than others. Many Census blocks have few or no people in them. Adding a little noise to large statewide measurements hardly makes a noticeable difference, of course. As originally designed, however, the original TopDown Algorithm defaults to equal settings for all six values of ε_1 through ε_6. That means that the variance of random noise injected at the state level is the same as the variance for noise injected to fuzz the population counts of Census blocks. Even if this policy is intuitive to those who care more about privacy, it is inconvenient for those who care more about accuracy. Thus, work is underway to test alternative algorithmic settings and implementations.

M4. Then there is still the matter of state population invariants. To handle precisely the case when the exact values of a function on a dataset are released alongside statistics to which noise has been added, Ligett et al. (2020) introduces the notion of "(ε, δ)-bounded leakage differential privacy" or blDP. This a relaxed variant of (ε, δ)-differential privacy as defined in the Appendix below, which in turn reduces to ε-differential privacy as defined above if $\delta = 0$. Not only does blDP still imply a corresponding version of privacy protection, but those guarantees also satisfy suitable generalizations of the Composition Theorem and the Post-Processing Theorem. The later result, like its original version, still requires that further computations do not refer back to the unprotected data, of course. The TopDown Algorithm violates this assumption in passing from the national to the state level because of its use of state population totals.

DRAFT

Raising this problem typically elicits one of three very different responses. The first is an ad hoc argument about how, apart from generating an accepted number for redistricting, no one really knows or cares very much about *precisely* how many people reside in a state. Adding noise or not therefore should matter very little. A second approach is to re-define what it means for a state dataset x' to be a neighbor of x relative to an invariant so that: a) x and x' must have the number of rows, one for each person in the state according to the publicly invariant count of the total population; and b) x and x' are identical in all their rows except one, where the attributes listed in that row may differ. Kifer and Machanavajjhala (2011) shows that the basic guarantees and theorems associated with DP go through with this understanding of neighborliness, though these results do need some re-interpretation in this context.

The third response suggests an alternative approach to fuzzing that avoids the need for such complicated post-processing altogether. The idea in Gong and Meng (2020) is not to use an off-the-shelf probability distribution to add randomness, but rather to condition that probability on the requirement that the privacy-protected results satisfy face-validity constraints and add up to the given invariant, for example. This is an elegant solution that makes statistical analysis much more straightforward and "congenial" in theory. It is not considered practical enough to implement officially in the 2020 Decennial, however, for at least two reasons: a) this would not provide output in a form compatible with the table-generating system that the Bureau is committed to re-using; and b) those conditional probability distributions are analytically intractable and so working with them, presumably by using a form of Markov Chain Monte Carlo sampling, would entail quite significant computational burdens. Note finally that, although we are illustrating how invariants pose challenges by discussing the basic P.L. 94-171 tables, the details are even more difficult in connection with the Demographic and Housing Characteristics (DHC) files.

For researchers wishing to explore such issues further, the Census has released Demonstration Data Products that consist of tables from the 2010 Census that have been processed using a preliminary version of the 2020 Disclosure Avoidance System. See https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html as well as https://nhgis.org/privacy-protected-demonstration-data.

IX. Ten Census Decisions Relevant to Researchers

Researchers therefore need to monitor the Census Bureau's decisions as it proceeds with implementing differential privacy for the 2020 Census data products. We summarize the likely impact on the main dissemination modes in Table 1.

DRAFT

Some of the changes are known and others have not been decided. Generally, compared to the 2010 Census, the impact of adopting differential privacy protections grows with the level of disaggregation. The Noisy Measurement File is new but access is uncertain. The DHC is also new and will look different and have different statistical properties than its predecessors, including the SF1. Special tabulations will likely be harder to come by than in the past. Microdata analysis will be available through the Federal Statistical Research Data Center system, possibly augmented via widely available synthetic data files.

As is apparent, the conduct and validity of research on 2020 Census data depends on key decisions that the Census will make going forward. Process choices of all sorts can still promise more or less accuracy. These decisions are hardly limited to ones about privacy-loss budgets. Timing and funding constraints also have dramatic effects on accuracy due to inevitable tradeoffs unrelated to privacy protection procedures, such as how much emphasis the Census effort as a whole places on advertising and on-line services, technology and security, canvassing and follow-up, de-duplication and data cleaning, the identification and imputation of missing data, etc. Because it is hard to quantify the cumulative consequences of these other policies, it has been convenient to forego close analysis of what differences they make. Now that we will be able to quantify the consequences of formal disclosure limitation procedures like differential privacy, we will be able to ask both how much accuracy we can really expect and how much privacy we want to give up in order to get it.

This transparency is new, challenging, and, for some long-time users of Census data and friends of the Census Bureau, disconcerting. Specifically with regard to how the concept of differential privacy will be applied in the 2020 Decennial, here are examples of important decisions that the community of researchers and other Census users should have opinions about that can be heard and taken into account. Most of these decisions will be made by the Census Data Stewardship Executive Committee, which charges the Census Disclosure Review Board with their enforcement (see https://www2.census.gov/foia/ds_policies/ds025.pdf).

D1. **The Value of ε^*.** This determines the total privacy-loss budget, of course. It is simple to explain to policymakers that smaller values of ε^* provide more privacy but less accuracy while larger values provide more accuracy but less privacy. More difficult is calibrating the impact of a given value of ε^* in comparison with more familiar measures of any sort, in order to inform the choice. Yes, we know what adding noise in the form of a random variable X with a known distribution whose mean is zero and whose standard of deviation σ is inversely proportional to epsilon (see Appendix). No simulations are necessary to discover that such a random variable can sometimes take on large values. Note that the total privacy-loss budget announced by the Census Bureau on June 9, 2021 for its redistricting data products, is larger than many observers feared. The increased privacy-loss budget over the levels reflected in the April 2021 demonstration data—which will lead to lower noise infusion than that in the April 2021 demonstration data—was primarily allocated to the total population and race by ethnicity queries at the block group level and above. (See https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-

DRAFT

management/process/disclosure-avoidance/2020-das-updates.html.)

But adding noise is not all that happens. Though not required to protect privacy, there are also the opaque adjustments made during post-processing that, for example, can bias small results away from zero in an effort to do away with counts that have gone negative after the addition of noise. The important question is not what noise injection does to the Census Edited File produced before any statistical disclosure limitation takes place, it is how the noise corresponding to a given privacy-loss budget compares to the other noise, miscounts, and inaccuracies—including both sampling and non-sampling errors—that are either already present in that Census Edited File or imposed afterwards by the post-processing procedures for providing publicly presentable tables. Without such estimates, how can anyone say whether setting $\varepsilon^* = 10$, for example, will be more like fiddling with the pennies when the dollar count is not even right or like paying attention to the dollars without bothering to count the change?

D2. **The Distribution of $\boldsymbol{\varepsilon^*}$ Along the Geographical Hierarchy** (ε_1 through ε_6) **and Among Other Attributes**. The Census Data Stewardship Executive Committee could benefit from guidance about choosing the portion of the privacy-loss budget allocated to different levels of geography or to other calculations. Dividing the budget equally among the levels, for example, would hit small areas with noise of the same variance as large ones. This is not likely to make everyone happy. Similar concerns apply to how the privacy-loss budget is allocated to protect particular attributes as opposed to others through, for example, the many choices that go into applying the Matrix Mechanism. Simulations based on public data from 2010 can help gauge the effects of such decisions on relative or absolute accuracy, for example. Note, however, that the 2010 data have been artificially homogenized by swapping. Current methods of disclosure avoidance based on differential privacy are more concerned with hiding outliers. Differing goals may therefore help account for observed differences in accuracy.

D3. **Invariants and Priorities within a Geographical Level.** The plan is for total U.S., state and territory populations to remain invariant, along with the number of housing units and the number and type of group quarters facilities at the block level. This means that their value will not change during the disclosure avoidance process. This could yet change as the Bureau proceeds. Since exact invariants blow the privacy-loss budget as it is usually understood, imposing approximate invariants is another option. Subject to the overall budget constraint, any given statistic or any given table can be more or less protected from noise. Which should be? More precisely, what considerations should go into the design of strategy matrices for implementing the Matrix Mechanism?

D4. **Small Subdivisions and Subgroups.** The wide variety of uses for various Census products range from redistricting to voting rights analysis and from rezoning to the triggering of various benefits for cities, tribes, or small regions. Geographical units

DRAFT

other than the state, county, and census block division that are part of the usual "spine" pose difficulties for any algorithm, not just the TopDown one. Even on the spine, over 1.8 million Census Blocks have only 1 to 9 residents whose privacy could easily be threatened by the release of too much information at that level. Some users would like accurate household data, too. The Census Bureau has so far resisted demands for granular data like this that would make it easy for analysts to recover, or pretend to recover, detailed information about individuals and families. As important as the consequences might be, decisions that depend on precise numbers like this can also be viewed as egregious *overfitting* to the data at hand. House-by-house redistricting as practiced by partisan consultants is just one example. Census tables are only based, after all, on one particularly comprehensive population sample rather than an ideal but impossible tally based on a current count of literally everyone. In addition to protecting confidentiality, query mechanisms that satisfy differential privacy also protect against overfitting in particular and p-hacking more generally (see Wasserstein and Lazar 2016). The reason is that output from such a mechanism cannot depend too sensitively on any one observation. For an example, see Dwork et al. (2015).

D5. **Post-Processing Criteria and Implementation**. Once fuzzed to protect privacy, tables will be subject to further processing to make them look more like Census products. That means eliminating negative numbers, imposing structural zeros, making sure columns sum properly, etc. There are many ways of finding tables that are close to the noisy ones but that look better. It all depends on what you mean by "close," for example. Most methods introduce statistical bias and other distortions of one kind or another. This has nothing to do with the trade-off between privacy and accuracy. It is instead trading away even more accuracy for the sake of formatting, face validity, and cosmetic considerations.

D6. **Researcher Access to Noisy Measurement Files (Pre-Post-Processed Data).** Even if the Census Bureau understandably balks at releasing tables with negative numbers and fractional counts to the public, researchers could benefit enormously from access to the noise-injected tables before they undergo post-processing. This would provably have no effect whatsoever on the privacy-loss budget. Failure to make the NMF available, however, would undermine the Census Bureau's claims of transparency about the new Disclosure Avoidance System. Since post-processing can impose distortions in unpredictable and unexplainable ways, it hardly matters if the NMF tables were constructed in explicit and explainable ways unless researchers have access to them. Even if the Bureau released the objective function and constraints used to produce the publicly released tables, the calculations still could not be readily reproduced or simulated by others.

D7. **Public Use Microdata Samples.** For previous Decennial Censuses, the Bureau has released 1% and 5% samples drawn from its records after stripping them of obvious identifiers. Some swapping and other suppression methods are said to be applied as

DRAFT

well. These files have been very useful for research and other purposes, as are ones similarly prepared based on samples from the American Community Survey (ACS). This approach to releasing data is incompatible, however, with the formal disclosure limitation guarantees offered by differential privacy. Ruggles (2019) notes that 60,000 or more individuals download over 100,000 Census and ACS IPUMS data files each year and other archives, such as the Inter-university Consortium for Political and Social Research, as well as the Census Bureau itself, serve hundreds of thousands of additional users. The Census Bureau has explicitly postponed any decision about changing privacy protection procedures for the ACS until at least 2025.

The Census Bureau has begun researching the potential release of privacy-protected microdata set that models the entire population. This would be a public file with microdata backed out from the official tables produced for Congress and the public. When tallied, the privacy-protected microdata would closely reproduce those privacy-protected tabulations as published. Such a file could be constructed and used freely with no further impact on the privacy-loss budget. The Census Bureau has already released a prototype of this kind of Privacy-Protected Microdata File (PPMF) based on the 2010 Census data. See https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-metrics.html).

Although such a PPMF will be, by design, entirely consistent with the published tables, there is no guarantee that correlations or other higher-order relationships computed from that microdata would necessarily track those in the confidential data set. Findings would need to be checked to make sure they are not spurious. Validation or verification servers can accomplish this by going back to the original data in ways that may incur only small and manageable costs in terms of the privacy-loss budget. For examples and explanations of academic research carried out this way, see Reiter (2019). For details about how the IRS is developing just such a system for facilitating research on its sensitive tax data, see Bowen et al. (2020).

D8. **Researcher Access to Confidential Microdata**. A growing share of academics analyze confidential microdata files in Federal Statistical Research Data Centers (FSRDCs). Working within highly protective procedures, policies, and spaces, qualified researchers pursue pre-approved projects on sensitive data. While the application process for this access can be challenging, The Foundations for Evidence-Based Policymaking Act (HR4174) recently passed by the Congress, implements a number of the recommendations regarding data access and privacy made by the Commission on Evidence-Based Policymaking, calls for the Office of Management and Budget (OMB). (See https://www.congress.gov/115/bills/hr4174/BILLS-115hr4174enr.pdf.) At this writing, work is underway to implement these provisions, including development of a common application process to access confidential Federal data.

The Census Bureau reviews all analytical output (tables, estimates, figures, etc.) for

DRAFT

compliance with standard privacy policies before researchers can take them out of an FSRDC. Now, the impact of any such releases on the privacy-loss budget will also need to be considered as well. With regard to standard differential privacy, making public the precise values of statistical queries blows that budget entirely. In any case, it would not be very practical to shift a large share of the research previously done on public use files to the FSRDCs since there are only 29 such facilities with a total of about 300 workstations among them. Current plans call for adding, on average, only three new sites per year.

D9. **Additional Tabulations and Other Decennial Census Products.** Along with the outputs already mentioned, the Census Bureau accepts requests to compile special tabulations for specific purposes. Prioritizing such requests will become all the more challenging since those that cannot be derived from data already released will make further demands on the privacy-loss budget. Anything requiring linkages with other datatsets is apt to be particularly problematic. On the upside, though, one advantage of application of differential privacy is that removes the need for thresholds, rounding, or cell suppression in special tabulations, which should make them more useful than in the past. Some users have therefore begun submitting their requests already in response to two Federal Register Notices that asked about high-priority use cases.

Note that all our discussions concern products based on the 2020 Decennial Census only. Approaches for other data programs that come under formal privacy protection must be tailored with respect to the type of data collection, the variables tabulated, the most essential use cases, and many other factors. So, while there is much to be learned from implementing differential privacy guarantees in the 2020 Census, data users should not assume that these same decisions will apply for other programs, such as the Current Population Survey or ACS. Abowd (2018) states, "The first Census Bureau product that will use the new system will be prototype redistricting data from the 2018 Census Test. This confidentiality protection system will provide the foundation for safeguarding all the data of the 2020 Census. It will then be adapted to protect publications from the ACS, economic censuses, and eventually all of our statistical releases." In 2019, the bureau made a commitment *not* to introduce differential privacy protection for the ACS until 2025 at the earliest. (See https://www.census.gov/newsroom/blogs/random-samplings/2019/07/boost-safeguards.html)

D10. **Relaxing the Privacy-Loss Budget Constraint over Time.** Absolute legislative prohibitions against the release of personally identifiable information collected by the 2020 Census will expire in 2092. Presumably, the risks and harms potentially caused by any such release subside over time rather than disappearing all at once at the end of 72 years. Does that mean that, over the years, the Census Bureau can gradually relax the privacy-loss budget determined when ε^* is initially set by the Data Stewardship Committee? This could have substantial impact on the planning and execution of research projects. How will such decisions be made? How can both research

advocates and privacy advocates not only have input, but also present evidence about the consequences of changes in epsilon? For example, Katharine Abraham (2019) suggests creating a "peer review" model similar to that used by the National Science Foundation or the National Institutes of Health to allocate research funding, with committees of scholars assessing the merits of competing proposals and their privacy-loss budget costs, could perhaps be adopted for this purpose."

Table 1 and this list of issues emphasize that there are enough changes ahead to 2020 Census products that users should re-examine any pre-existing inclinations or decisions about optimal access modes.

X. Conclusion

Benjamin Franklin once said, "By failing to prepare, you are preparing to fail." While the Census Bureau has not yet released a full disclosure plan for the 2020 Census, they have been open about the approach they will take and how it is based on ideas about differential privacy (Abowd et al. 2020). This paper has focused on known implications for researchers.

In this new world, researchers planning to use 2020 Census data will face a reorganized array of access options, each with a usefulness that depend on project goals as well as decisions made by the Census Bureau. For some purposes, published or special tabulations may be more useful than before, and microdata less useful. For other purposes, the reverse could be true. Furthermore, the merits of every "consultation" with the confidential microdata will be judged against its privacy loss to determine if it can be approved.

First, we summarize some research-friendly recommendations for the Census Bureau.

R1. **Researchers need access to the Noisy Measurement File for the 2020 Census.** We strongly recommend that the Census Bureau release the Noisy Measurement File, not solely post-processed files. Releasing the noisy file entails no additional privacy loss, so the Bureau should make it broadly available for research purposes, despite the communications and resource challenges involved. Such release needs to include adequate documentation and an efficient access procedure.

R2. **The Bureau should prioritize expanding and speeding researchers' secure access to 2020 Census data through the FSRDC system.** Without major changes to accessibility, constraints on physical as well as review capacities may therefore prevent many otherwise worthy proposals from going forward in a timely manner. Other countries have established secure remote access to their equivalent facilities. In addition, the Bureau will need a process to manage the privacy budget implications for validation of results obtained from behind-the-firewall research.

DRAFT

R3. **For broad use, the release of Privacy-Protected Microdata Files (PPMF) should be accompanied by a validation process.** Census must support research conducted on a such microdata files with a means to validate the resulting inferences. Each verification will incur modest costs to the privacy-loss budget. Without this capability, the other way to check findings would be to test programs on the PPMF that would eventually be run on unprotected microdata within the safety of the (overtaxed) FSRDC system. As always, the release of calculations run on original data need fuzzing that again draws on the privacy-loss budget.

This brings us back to the question posed in our title. What do researchers need to know? Here, in summary, are six points for users of Decennial Census data to keep in mind:

P1. **Researchers should analyze the Noisy Measurement File whenever practical.** The TopDown procedure introduces random noise to aggregations to produce a noisy file (that is, with no post-processing applied). The "post-processing" to make the data look better would be used as planned in the final public release. Most researchers should rely on the Noisy Measurement File, not post-processed files, because its properties will be far more certain and, thus, easier to account for statistically. As statistical agencies move to adopt formal privacy protection for other programs, we believe that this approach will likely generalize to products beyond the 2020 Census.

P2. **Researchers should update their usual analytical toolboxes to account for fuzzed data.** For example, they need to take seriously the variance addition rule. Noise may "average out" when you take averages, but it adds when you do addition. This fact lies behind the bureau's decision to go with the TopDown approach. More generally, using privacy-protected data properly will require researchers to use analytical techniques that explicitly treat the Decennial Census as a noisy sample of the U.S. population rather than an actual count of the entire U.S. population. The organization of research, not just on Census data but empirical research generally, will need to adjust to take these realizations into account. Analysts will need to develop approaches that can reduce leakages, such as working with "backed-out" data and robust estimators. Finally, researchers need to learn how to prioritize research questions. They typically regard data as a non-rival commodity, that is, a good whose consumption by one party does not reduce the ability of another party to use it as well. Evidence, however, is rival—at least when viewed from within the conceptual framework that goes along with formal disclosure limitation methods. For a given dataset containing confidential information, *every query answered inevitably leaks some privacy*. Conducting research using only query mechanisms that satisfy ε-differential privacy slows and calibrates the rate of leakage.

P3. **Researchers should recognize that differential privacy protections convey the additional benefit of deterring overfitting.** Overfitting and p-hacking are serious problems in the conduct of empirical research on sampled data (see Wasserstein

DRAFT

2016). In both cases, the mistake is presenting conclusions that depend on the particular sample as if they hold for the entire population. In addition to leaking privacy, *every query answered also leaks some validity* because each answer further facilitates more overfitting and p-hacking. Conducting research using only query mechanisms that satisfy ε-differential privacy slows and calibrates the rate of leakage. This holds even when dealing with data about stars or fish where confidentiality is not an issue. The reason, of course, is that such techniques prevent statistical answers from depending too much on the presence or absence of one particular observation.

P4. **Researchers should remember, and remind others, that the purpose of these changes is to protect respondents' privacy so that decision-makers, academics, and the public at large will continue to have access to valuable information.** One of the greatest threats to the accuracy of Census data is non-participation. Many residents may feel that the risks of providing sensitive information about themselves are too great unless their privacy is securely protected. The Census Bureau therefore has a statutory obligation to protect the privacy of respondents and is subject to Congressional oversight. Were the Bureau to fail to provide state-of-the-art privacy protection, decision-makers outside the statistical agency (in Commerce, Congress, or the courts) might well impose another privacy protection regime that could severely curtail research access and/or lack many of the desirable features of differential privacy protections, such as transparency. The growth of computing power and other data sources have heightened privacy concerns among citizens. Inadequate privacy protections could therefore suppress response rates in the Census and other federal surveys, exacerbating their decades-long decline.

P5. **Without delay, researchers should seek to provide more input into the many decisions the Census Bureau needs to make to fully implement differential privacy.** Subject to the legal, mathematical, and practical constraints sketched here, the Census Bureau needs structured and purposeful community input in its work to produce Decennial Census products that will be both safe and useful. Academics and social scientists can and should try to understand and influence the details of those implementations. One route is by responding to Federal Register Notices, contacts at Census (including the email address 2020DAS@Census.gov that forwards to key staff working on DAS in the Bureau) and through their contacts at Census and on the Census Bureau's advisory committees. (See https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html. For information on the members, meetings and agenda for Census advisory committees, see https://www.census.gov/about/cac.html. Meetings are open to the public. Another avenue is through the National Academies of Sciences, Engineering and Medicine's Committee on National Statistics (CNSTAT), which the Census Bureau has commissioned to hold workshops to provide input into these decisions. CNSTAT held a "Workshop on 2020 Census Data Products: Data

DRAFT

Needs and Privacy Considerations" on December 11-12, 2019. At least one more is planned. For information, see https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations. A third way is through professional associations such as the American Economic Association, the American Statistics Association and the American Population Association. Such feedback is essential because it is naïve to assume that Census staff will be familiar with all considerations of importance to researchers otherwise.

P6. **Finally, researchers should actively help preserve the quality of the Decennial Census and other official statistics.** Delineating the trade-off between privacy and accuracy—as formal disclosure limitations methods do—highlights the need for full participation in the Decennial Census and federal surveys in general. This starts with confidence among potential respondents in the basic procedures and protections implemented by the Bureau and other statistical agencies on an ongoing basis. No amount of noise reduction can make up for an undercount. If the Bureau failed to protect privacy now and Census data were weaponized, the accuracy of subsequent Censuses would surely suffer. Then, perfect accuracy now with respect to the actual responses now would come at the cost of far less accuracy in the future. Everyone who cares about basing good research on data produced by the federal statistical system data in the future should help promote trust and full participation in officially approved federal surveys of all kinds. In addition to its legal mandates, this is another key reason why the Census Bureau is, carefully but with determination, working to find the best methods for both protecting respondents' privacy and providing useful statistics.

DRAFT

Table 1
Comparison of access options for research using 2010 and 2020 Decennial Census data

	2010 Census disclosure	2020 Census disclosure plan
Governance	• Plan set by Data Stewardship Executive Policy Committee (DSEP). • Releases approved by Census Disclosure Review Board.	
Apportionment tabulations (state level)	• State counts as enumerated not included in disclosure processes. Aggregation judged sufficient to protect privacy.	
Noisy Measurement File(s)	• Not applicable.	• No decision made to date concerning release. • Privacy-protecting noise added with no post-processing. • No privacy loss by release.
Redistricting tabulations	• Swapping and nontransparent privacy-protecting noise added to results for smaller and/or random blocks.	• Predesignated "invariants" (as defined by DSEP) reflect enumerated totals. • Privacy-protecting noise with announced properties applied to other aggregations, with adding up likely to be preserved. Large aggregations that are not invariants will reflect minimal noise, by design.
Balance of 2010 Decennial Data Products (Demographic Profiles, SF1, SF2, Congressional Districts Summary Files, American Indian and Alaskan Native files)	• Nontransparent methods preserve privacy (tabulations based on microdata protected by swapping to produce Hundred Percent Detail File (HDF).	• See Census Bureau's "crosswalk file" for planned correspondence between the 2010 Demographic Profiles and 2020 Demographic Profiles, as well as between the 2010 SF1 and the 2020 Demographic and Housing Characteristics (DHC) File. Contents of other files have not been determined. • Large aggregations and predesignated "invariants" (defined by DSEP) reflect enumerated totals. • Privacy-protecting noise with announced properties applied to other aggregations. No decision yet made about whether adding up will be preserved. • Less geographic detail than in 2010.
Additional special tabulations for federal, state, local and other data users	• Nontransparent methods preserve privacy by rounding, thresholds and cell suppression.	• Subject to transparent overall privacy-loss budget and priorities determined by Census. • All tabulations charged against overall privacy-loss budget. • Privacy protected by transparent addition of noise. No decisions made to date on adding up.

DRAFT

		• No decisions made to date on cell suppression. Purpose would be meet fitness for use criteria, not to further protect privacy.
Public use microdata files (PUMS)	• Only samples released. • Privacy protected by nontransparent swapping, deletion of fields, and other methods.	• Synthetic data under consideration, but no decisions made to date.
Confidential microdata analysis at Federal Statistical Research Data Centers	• Approved projects use HDF (after swapping and other privacy protections are applied) in a highly protected environment. • Analytical output reviewed for compliance with privacy protections.	• No change in process or capacity. • Subject to transparent overall privacy-loss budget and priorities determined DSEP. • Output for each project charged against overall privacy-loss budget.

Notes:

- DSEP approved this final list of invariants on November 24, 2020: state population totals, the number of housing units at the block level, and the number and type of group quarters facilities at the block level. See https://content.govdelivery.com/accounts/USCENSUS/bulletins/2ae5eda.
- 2020 Demographic and Housing Characteristics (DHC) File replaces the 2010 SFI and SF2 files.
    - 2010 Summary File 1 (SF1) contains the data compiled from the questions asked of all people and about every housing unit. Population items include sex, age, race, Hispanic or Latino origin, household relationship, household type, household size, family type, family size, and group quarters. Housing items include occupancy status, vacancy status, and tenure (whether a housing unit is owner-occupied or renter-occupied). See https://www.census.gov/prod/cen2010/doc/sf1.pdf.
    - 2010 Summary File 2 (SF 2) contains the data compiled from the questions asked of all people and about every housing unit. SF 2 includes population characteristics, such as sex, age, average household size, household type, and relationship to householder such as nonrelative or child. The file includes housing characteristics, such as tenure (whether a housing unit is owner-occupied or renter-occupied), age of householder, and household size for occupied housing units. Selected aggregates and medians also are provided. See https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/complete-tech-docs/summary-file/sf2.pdf.

DRAFT

References

Supreme Court decisions

DEPARTMENT OF COMMERCE et al. v. UNITED STATES HOUSE OF REPRESENTATIVES et al., No. 98–404. Argued November 30, 1998—Decided January 25, 1999* https://supreme.justia.com/cases/federal/us/525/316/

UTAH V. EVANS (01-714) 536 U.S. 452 (2002) 182 F. Supp. 2d 1165, affirmed.

Abowd, J.M., 2018a, July. "The US Census Bureau adopts differential privacy." In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2867-2867). ACM.

Abowd, J.M., 2018b, July. "Staring-Down the Database Reconstruction Theorem." In Joint Statistical Meetings, Vancouver, BC.

Abowd, J., Ashmead, R., Simson, G., Kifer, D., Leclerc, P., Machanavajjhala, A. and Sexton, W., 2019. Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge. *US Census Bureau*.

Abowd, John M., Gary L. Benedetto , Simson L. Garfinkel , Scot A. Dahl, Aref N. Dajani, Matthew Graham, Michael B. Hawes, Vishesh Karwa, Daniel Kifer, Hang Kim, Philip Leclerc, Ashwin Machanavajjhala, Jerome P. Reiter, Rolando Rodriguez, Ian M. Schmutte, William N. Sexton, Phyllis E. Singer, and Lars Vilhuber. 2020. "The modernization of statistical disclosure limitation at the U.S. Census Bureau." https://www.census.gov/library/working-papers/2020/adrm/modernization-statistical-disclosure-limitation.html

Abraham, Katharine, 2019. "Reconciling Access and Privacy: Building a Sustainable Model for the Future." Unpublished working paper, University of Maryland, January 2019.

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F. and Talwar, K., 2007, June. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 273-282). ACM.

Bowen, C.M., Bryant, V., Burman, L., Khitatrakun, S., McClelland, R., Stallworth, P., Ueyama, K. and Williams, A.R., 2020, September. A synthetic supplemental public use file of low-income information return data: methodology, utility, and privacy implications. In *International Conference on Privacy in Statistical Databases* (pp. 257-270). Springer, Cham.

Data, T., Duncan, G.T., Fienberg, S.E. and Krishnan, R., 2001. "Confidentiality, and data access: Theory and practical applications for statistical agencies."

Dinur, I. and Nissim, K., 2003, June. "Revealing information while preserving privacy." In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 202-210). ACM.

Dwork, C., McSherry, F., Nissim, K. and Smith, A., 2006, March. "Calibrating noise to sensitivity in private data analysis." In Theory of cryptography conference (pp. 265-284). Springer, Berlin, Heidelberg.

Dwork, C. and Roth, A., 2014. "The algorithmic foundations of differential privacy." Foundations and Trends® in Theoretical Computer Science, 9(3–4), pp.211-407. (See especially Chapter 8).

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O. and Roth, A., 2015. "The reusable holdout: Preserving validity in adaptive data analysis." Science, 349(6248), pp.636-638.

Evans, G., King, G., Schwenzfeier, M. and Thakurta, A., 2019. Statistically valid inferences from privacy protected data. https://gking.harvard.edu/dpd2

Evans, G. and King, G., 2020. Statistically valid inferences from differentially private data releases, with application to the facebook urls dataset. *Political Analysis. URL: GaryKing. org/dpd*.

Federal Committee on Statistical Methodology, 2005. "Report on Statistical Disclosure Limitation Methodology." STATISTICAL POLICY WORKING PAPER 22 (Second version, 2005.) https://www.hhs.gov/sites/default/files/spwp22.pdf

Gong, R. and Meng, X.L., 2020, October. Congenial Differential Privacy under Mandated Disclosure. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference* (pp. 59-70).

Groves, R.M. and Lyberg, L., 2010. Total survey error: Past, present, and future. *Public opinion quarterly*, *74*(5), pp.849-879.

Kifer, D. and Machanavajjhala, A., 2011, June. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data* (pp. 193-204).

Kinney, S.K., Karr, A.F. and Gonzalez Jr, J.F., 2010. Data confidentiality: the next five years summary and guide to papers. *Journal of Privacy and Confidentiality*, *1*(2).

Kukutai, T., Thompson, V. and McMillan, R., 2015. "Whither the Census? Continuity and change in Census methodologies worldwide, 1985–2014." *Journal of Population Research*, *32*(1), pp.3-22.

Li, C., Miklau, G., Hay, M., McGregor, A. and Rastogi, V., 2015. "The matrix mechanism: optimizing linear counting queries under differential privacy." *The VLDB journal*, *24*(6), pp.757-781.

Ligett, K., Peale, C. and Reingold, O., 2020. Bounded-leakage differential privacy. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

McGeeney, Kyley, Brian Kriz, Shawnna Mullenax, Laura Kail, Gina Walejko, Monica Vines, Nancy Bates, and Yazmín García Trejo, 2019. "2020 Census Barriers, Attitudes, and Motivators Study Survey Report: A New Design for the 21st Century." US Census Bureau, January 24, 2019. https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-report-cbams-study-survey.pdf

National Academies of Sciences, Engineering, and Medicine. 2021. Principles and Practices for a Federal Statistical Agency, Seventh Edition. Washington, DC: The National Academies Press. https://doi.org/10.17226/25885

DRAFT

Porter, T.M., 1996. *Trust in numbers: The pursuit of objectivity in science and public life.* Princeton University Press.

Reamer, A., 2020. "Counting for dollars 2020: The role of the Decennial Census in the geographic distribution of federal funds." https://gwipp.gwu.edu/counting-dollars-2020-role-decennial-census-geographic-distribution-federal-funds

Ruggles, Steven, 2019. "Implications of Differential Privacy for Census Bureau Data and Research," University of Minnesota. January 2019.

Wasserstein, Ronald L. and Nicole A. Lazar. 2016. "The ASA Statement on p-Values: Context, Process, and Purpose." *The American Statistician*. Volume 70, 2016 - Issue 2 https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108

DRAFT

Appendix:

Primer on Differential Privacy and How It Works

For researchers new to working with data with Differential Privacy protections, this appendix introduces a few basics including: 1) what situations the concept applies to; 2) what the definition means, both in words and in simple mathematical terms; 3) how to implement a query mechanism to protect privacy; 4) consequences of the definition that explain what privacy budgets are and how they get used up; and 5) some practical notes about variants and applications.

1. The Concept

Motivation for the idea of Differential Privacy derives from the following situation. An authority is collecting data about individuals for research purposes. Suppose a potential participant in this study is concerned about protecting his or her privacy. What rigorous guarantees could reassure this person about the safety of allowing his or her personal information to be included in the dataset that researchers will analyze?

One worry, of course, could be a data breach. Here we put that concern aside to concentrate on privacy rather than cybersecurity.

Specifically, we imagine that a trusted curator provides the only possible access to the dataset x. A researcher wishing to analyze the dataset can submit her question to the curator, who runs a "query mechanism" denoted $\mathcal{M}$ that calculates the response $\mathcal{M}(x)$. The extent to which privacy will be protected depends on the properties on the mechanism $\mathcal{M}$. If, for example, the question submitted asks for the social security number of the person whose information appears in line 17 of the dataset, a curator concerned with privacy should not implement a mechanism that simply reports that entry. Otherwise, an analyst could easily tell that the person with the social security number returned did contribute data. For a study of some rare or embarrassing disease, that finding alone could significantly compromise a participant's privacy. Perhaps $\mathcal{M}$ should only supply a few digits of the requested social security number, or perhaps some of the digits should be randomly scrambled or slightly altered before presenting $\mathcal{M}(x)$ back to the analyst?

In fact, the whole idea of Differential Privacy is to limit quite formally how much an analyst seeking to discover who was or wasn't listed in dataset x can learn from the query response $\mathcal{M}(x)$. After all, if an analyst cannot even tell whether my information is included in the dataset, then she cannot find out anything about me personally because of my participation in the study. In that case, each individual could donate personal data with utter confidence that doing so will not violate his or her privacy at all.

DRAFT

This would be the ideal, anyway. Both in practice and in theory, however, there is an unavoidable trade-off between the accuracy with which $\mathcal{M}(x)$ answers the analyst's original question and the risk of a privacy violation. Everyone who has ever tried to protect the personal information in a dataset knows that such a trade-off exists. The usual approach is to suppress or garble data details that could obviously identify people, and then hope for the best. In contrast, the concept known as Differential Privacy provides a rigorous way of understanding and controlling that trade-off between accuracy and privacy in terms of a specific parameter denoted by ε, the Greek letter epsilon that mathematicians typically use to stand for a small positive number.

2. The Definition

Given a query mechanism $\mathcal{M}$ defined on a collection of datasets X, let's define what it means for $\mathcal{M}$ to satisfy "ε-differential privacy." We are imagining an analyst who submits a question to the data curator. Perhaps the analyst wants to know the value of some function f when applied to the dataset x that the curator keeps securely behind a firewall. For example, $f(x)$ might be the average of a particular column in the spreadsheet called x. Based on the analyst's query, the curator calculates $\mathcal{M}(x)$ and returns information about that value in response. A responsible curator does not necessarily make $f(x)$ and $\mathcal{M}(x)$ equal. Rather, the mechanism may involve some randomization so as not to reveal too much about the confidential information in the dataset, or even about which dataset x in X the curator has at hand.

Based on the information $\mathcal{M}(x)$ that she receives, the analyst revises her statistical beliefs about the dataset x. Suppose that I am a potential participant in the study who is concerned about the privacy risks of donating personal information. To decide whether to take that risk, I should worry about how much $\mathcal{M}$ can reveal to an analyst about whether or not my information is even listed in the dataset x at all. Specifically, the question I need to answer is how well the query mechanism $\mathcal{M}$ can help an analyst distinguish whether $x = d$, where d is a dataset that has my personal information in one of its rows, or $x = d'$, where d' is a dataset that is identical to d except for that one row which is altered or missing. Two datasets in X that differ by exactly one row like this are called "neighbors."

Good analysts revise their beliefs using methods first devised by the Reverend Bayes in the mid-1700's. He imagined an analyst who wants to adjust the probability of an event A given that a random event B has actually occurred. Let $\Pr(A)$ and $\Pr(B)$ denote the probabilities assigned to those events before the news arrives that B has occurred. The "conditional probability of A given B" is defined as:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

This is readily interpreted as the fraction of the times when B occurs that A does as well. (Note that $\Pr(A|B)$ does not equal $\Pr(B|A)$ in general, and that many mistakes result from thinking

DRAFT

otherwise.) Rearranging this definition and cancelling some denominators gives the following form of "Bayes' Law":

$$\frac{\Pr(A|B)}{\Pr(A'|B)} = \frac{\Pr(B|A)}{\Pr(B|A')} \times \frac{\Pr(A)}{\Pr(A')}$$

In this remarkably useful formula, the ratio of probabilities on the far right is called the "prior odds" of A versus A'. The ratio on the left is called the "posterior odds" of A versus A' given B. The ratio in the middle is called the "Bayes Factor". The closer it is to one, clearly the less an analyst learns about A versus A' by finding out B.

In our situation, the analyst assigns her own personal probabilities to the events $x = d$ and $x = d'$ before learning anything about $\mathcal{M}(x)$. Suppose the event B that the curator tells her has occurred is $\mathcal{M}(x) \in S$. In other words, she finds out that $\mathcal{M}(x)$ belongs to some subset S in the range of values $\mathcal{M}$ can take on over X. Substituting into Bayes' Law yields:

$$\frac{\Pr(x = d \mid \mathcal{M}(x) \in S)}{\Pr(x = d' \mid \mathcal{M}(x) \in S)} = \frac{\Pr(\mathcal{M}(x) \in S \mid x = d)}{\Pr(\mathcal{M}(x) \in S \mid x = d')} \times \frac{\Pr(x = d)}{\Pr(x = d')}$$

where the probability $\Pr(E)$ of an event E is taken jointly over the analyst's beliefs and over the randomness of the query mechanism $\mathcal{M}$. The Bayes Factor in the middle, however, simplifies to:

$$\frac{\Pr(\mathcal{M}(d) \in S)}{\Pr(\mathcal{M}(d') \in S)}$$

with the probability taken only over the randomness of the mechanism. As noted above, the closer this ratio is to one, the less well $\mathcal{M}$ can help an analyst distinguish between neighboring datasets, and hence the better $\mathcal{M}$ protects my privacy.

One convenient way to measure how close a number k is to one is by finding an $\varepsilon > 0$ such that $\exp(-\varepsilon) \leq k \leq \exp(\varepsilon)$ since, for small ε, $\exp(\varepsilon) \approx 1 + \varepsilon$ if we ignore higher order terms. Applying this to the Bayes Factor, we say that a query mechanism $\mathcal{M}$ defined on X satisfies "ε-differential privacy" if, for all neighboring datasets d and d' in X and for all sets S in the range of $\mathcal{M}$, we have:

$$\frac{\Pr(\mathcal{M}(d) \in S)}{\Pr(\mathcal{M}(d') \in S)} \leq \exp(\varepsilon)$$

Note that bounding the ratio from below is taken care of by reversing the roles of d and d'. To avoid questions about how to interpret this when the denominator could be zero, this criterion is usually expressed as:

$$\Pr(\mathcal{M}(d) \in S) \leq \exp(\varepsilon) \Pr(\mathcal{M}(d') \in S). \qquad (*)$$

DRAFT

For small ε, this definition means that, upon learning that $\mathcal{M}(x) \in S$ from the curator of dataset x, an analyst's prior odds about whether my information is even listed in x can only change by a factor that differs from one by about ε.

In other words, the smaller the ε, the less an analyst can learn about individuals from a query mechanism that satisfies ε-differential privacy. Intuitively, ε measures a kind of "plausible deniability." That is, if you want to claim that your data is not even in the dataset at all, ε controls the likelihood that an analyst could verify or falsify your claim based on a query made with a mechanism satisfying ε-differential privacy.

This does not, of course, prevent my being harmed by statistical findings calculated from a given dataset. To take a familiar example, statistical studies on a health outcome dataset should be able to test whether smoking causes cancer. Then, based on knowledge that I am a smoker from other sources, my health insurance company could decide to raise my premium rates. But as long as the calculations are carried out by a query mechanism satisfying ε-differential privacy, the research findings about cancer would hardly have been any different regardless of whether my information was in the dataset or not. That is, my insurance company should not be able to glean much actionable information about me as an individual from privacy-protected answers to queries made about the dataset. Differentially private mechanisms protect very precisely against "participation risk" by limiting any harm to me attributable to the presence or absence of my data in the dataset under study. When presented with a survey to fill out or a privacy agreement to sign, this is the kind of risk that should concern a potential respondent.

3. Implementing a Query Mechanism

How does this work in practice? Having developed a conceptual framework for discussing data privacy, we can now consider approaches a data curator could actually use to satisfy the definition of ε-differential privacy. Clearly, the curator cannot simply give an analyst precise answers to statistical questions—even innocent looking ones like averages or percentiles. Otherwise, it would be easy to find an S such that the left side of the Differential Privacy criteria (*) above equals one but the right side is zero, making the inequality impossible to satisfy for any ε.

So to protect privacy when an analyst asks for an average, a count, or some other statistic $f(x)$, the curator can calculate it behind the firewall but cannot deliver $f(x)$ untouched. A responsible curator will instead add a small amount of random noise to $f(x)$ before returning the sum, $\mathcal{M}(x)$, in answer to such a query. The parameter ε governs the inevitable trade-off between privacy and accuracy. More noise, and hence more privacy protection, goes along with smaller ε, but that also provides less accuracy.

DRAFT

What do we mean by "more noise"? The added random variable can be drawn from a distribution that is highly concentrated around zero (to minimize distortions), that has mean zero (to avoid introducing bias), and that has variance inversely proportional to ε^2 (to properly protect privacy). So, you can think of adding noise as a way of saying adding variance.

Specifically, suppose $f(x)$ is a statistic calculated from a database x. Consider a query mechanism of the form $\mathcal{M}(x) = f(x) + r$ where r is a random variable. We call this the "Laplace mechanism", for example, if the distribution of r is Laplacian with mean zero and hence density that decreases exponentially with distance from the origin. The important theorem is that such a mechanism satisfies ε-differential privacy as long as the Laplacian distribution has variance $\sigma^2 = 2(\Delta f)^2\varepsilon^{-2}$ where Δf stands for the "sensitivity of f." This is defined as the maximum of $|f(x) - f(x')|$ over all pairs of neighboring datasets x and x' in X. For counting queries of the sort that go into making most Census tables, it is easy to see $\Delta f = 1$ since adding or deleting a single individual from a dataset cannot change the value of f by more than that. For other statistics, such as regression coefficients, Δf may be quite difficult to bound due to the familiar sensitivity of such estimates to the inclusion or absence of data points that are "outliers."

We end this section with three notes.

- The appropriate value of ε for a given situation is not a mathematical question, but rather *a policy decision* for the curator to make and maintain.
- Protecting privacy this way is not necessarily about blanketing the entries in the original dataset with random perturbations, but rather about 'fuzzing' specific statistics calculated from that data before they are released to the analyst.
- More parenthetically, we use the terms "fuzzy," "fuzzed," and "unfuzzed" to refer to intentionally injecting random noise into data in order to preserve privacy. Although these are not common or technical terms in the literature, we find them appropriate and useful.

4. Properties of Query Mechanisms and the Privacy-Loss Budget

Two mathematical results bear on the challenge of selecting and respecting epsilon. One simplifies matters and the other complicates them.

First, the simplifying result. According to the "Post-Processing Theorem," differential privacy guarantees are immune to future threats—such as the unforeseen appearance of new computing or data resources. Specifically, having answered a single query though a mechanism satisfying ε-differential privacy can never increase the odds of determining whether an individual's confidential data belongs to the given dataset by more than a factor of $exp(\varepsilon) \approx 1 + \varepsilon$ no matter what else anyone ever does or reveals without accessing the original data again. The ability of one such query answer to make any difference is forever limited.

DRAFT

To see why the Post-Processing Theorem holds, let G denote a possibly random function on the range of a query mechanism $\mathcal{M}$ that satisfies "ε-differential privacy." Consider the composition query mechanism $G \circ \mathcal{M}$ defined by setting $G \circ \mathcal{M}(x) = G(\mathcal{M}(x))$ for all x in X. Suppose further that G is independent of $\mathcal{M}$ and x. For T a subset of the range of G, we have:

$$\Pr(G \circ \mathcal{M}\ (x) \in T) = \Pr(\ \mathcal{M}(x) \in G^{-1}\ (T)) = \Pr(\mathcal{M}(x) \in S) \times \Pr(S = G^{-1}(T))$$

where the probability is jointly over the independent randomness of G and $\mathcal{M}$ in the first two expressions, then over the appropriate marginals in the third. Taking $S = G^{-1}(T)$ in the definition of ε-differential privacy for $\mathcal{M}$ shows that $G \circ \mathcal{M}$ satisfies the same condition with the same ε.

In this argument, G does not depend on x. But suppose that, after learning $\mathcal{M}_1(x)$, the analyst decides to submit another question defined on X that the curator answers by calculating $\mathcal{M}_2(x)$. What happens to differential privacy guarantees after multiple queries? Unlike informal privacy protections that provide no guarantees whatsoever in such cases, differential privacy guarantees add up nicely when invoked repeatedly. To wit, assume a dataset curator first uses a query mechanism $\mathcal{M}_1$ that satisfies ε_1-differential privacy and uses $\mathcal{M}_2$ that satisfies ε_2-differential privacy. As long as the randomness of these two mechanisms are independent, then the combined mechanism $(\mathcal{M}_1, \mathcal{M}_2)$ that answers both satisfies $(\varepsilon_1 + \varepsilon_2)$-differential privacy. This "Sequential Composition Theorem" follows immediately from noticing that:

$$\frac{\Pr(\mathcal{M}_1(d) \in S, \mathcal{M}_2(d) \in T)}{\Pr(\mathcal{M}_1(d') \in S, \mathcal{M}_2(d') \in T)} = \frac{\Pr(\mathcal{M}_1(d) \in S)}{\Pr(\mathcal{M}_1(d') \in S)} \times \frac{\Pr(\mathcal{M}_2(d) \in T)}{\Pr(\mathcal{M}_2(d') \in T)}$$

$$\leq \exp(\varepsilon_1)\exp(\varepsilon_2) = \exp(\varepsilon_1 + \varepsilon_2)$$

where again the probability is joint on the left and then marginal on the right.

The Sequential Composition Theorem makes accounting for multiple queries easy in principle, but it poses a complicated bookkeeping challenge for curators charged with trying to maintain privacy guarantees over time. Suppose that the curator's policy is to provide privacy protections that are always at least as good as those corresponding to ε^*. If presented with a series of n queries from mechanisms satisfying ε_i-differential privacy for $i = 1, 2, \ldots, n$, the curator must make sure that $\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_n \leq \varepsilon^*$. In other words, the choice of ε^* imposes a *privacy-loss budget* that gets consumed as curators answer questions submitted by analysts. Answering too many such queries too accurately could exceed ε^*. Before that happens, a curator committed to maintaining ε^* has to stop entertaining queries altogether, shutting down further analyses that depend on new queries.

To reiterate, a curator charged with preserving a fixed ε^* privacy protection threshold in perpetuity faces a complex task. The curator must assign an ε_i for each query on an ongoing basis, mindful of the past and the future. That series of epsilons explicitly balances all privacy losses from current and future queries against one another. Highly accurate answers early on

DRAFT

will limit accuracy in future queries, or even prevent answering entirely. On the other hand, very inaccurate responses early on may impede statistical inference unnecessarily, perhaps because interest in the data fade over time. They could leave too much privacy on the table.

Note that the one special case in which each query does not necessarily use up more privacy is when the different queries deal with disjoint subsets of the population. Think of these as bins or buckets that are exclusive but collectively exhaustive. In that case, the "Parallel Composition Theorem" says that the total cost to the privacy-loss budget is the maximum rather than the sum of all the epsilons. This makes sense because, under such assumptions, an analyst's probability of finding me in the dataset as a whole is the same as her probability of finding me in one of those disjoint segments of the population.

5. Applications and Variants

Implementing Differential Privacy protection in a particular setting can be quite subtle and bespoke, depending both on the nature of both the datasets and on the nature of the research questions of interest. We mention a few dimensions of variation and their causes below.

First, certain query mechanisms defined on specific domains either do or do not satisfy ε-differential privacy. That is, the Laplace Mechanism that we highlight above is the most important one, but there are others. The Geometric Mechanism, for example, is a discrete version of the Laplace Mechanism.

In addition, policy trade-offs will vary between applications and instances. Since the community of researchers usually has more than one query to submit, the process of setting, allocating and managing the privacy-loss budget is a major policy challenge that will lead to variation in experiences. Mathematics can explain the meaning and consequences of such choices, but making those trade-offs involves values and not just technicalities.

Another challenge in some contexts can be defining an appropriate notion of when two datasets should be considered neighbors. This is especially the case when dealing with records of people interacting over networks, for example. This is one of many reasons why the basic treatment of Differential Privacy presented here has inspired many refinements, extensions, relaxations, and variants. These range from "local differential privacy" to "federated differential privacy" and from "bounded differential privacy" to "unbounded differential privacy."

Particular useful can be the slightly relaxed criteria whereby a query mechanism is said to satisfy (ε, δ)-differential privacy if, for all neighboring datasets d and d' in X and for all sets S in the range of $\mathcal{M}$, we have:

$$\Pr(\mathcal{M}(d) \in S) \leq \exp(\varepsilon) \Pr(\mathcal{M}(d') \in S) + \delta.$$

DRAFT

This is also known as "approximate differential privacy." A suggestive interpretation of this definition is that it requires ε-differential privacy with probability $1 - \delta$. The Gaussian mechanism, so called because it adds noise sampled from a normal distribution, notably satisfies (ε, δ)-differential privacy but not ε-differential privacy.

The Census Bureau is not only experimenting with implementation of (ε, δ)-differential privacy, but also with "concentrated differential privacy" as well. By imposing restrictions on the distribution of privacy losses rather than absolute bounds on those values, this other relaxation of the standard definition implies accuracy improvements, better group privacy properties, and tighter estimates of expectations while still preserving the Composition Law and other key properties of ε-differential privacy.

Finally, there are also software and hardware challenges when implementing any of these ideas. Ideally, code from protecting privacy cannot be too private since open source software is more readily checked for inadvertent errors or unrecognized withdrawals from the privacy-loss budget.

DRAFT