UNITED STATES DISTRICT COURT
FOR THE MIDDLE DISTRICT OF ALABAMA
EASTERN DIVISION

THE STATE OF ALABAMA, et al.,	)	
	)	
Plaintiffs,	)	
	)	
v.	)	**Case No.:**
	)	**2:21-cv-00211-RAH-ECM-KCN**
	)	
UNITED STATES DEPARTMENT OF COMMERCE, et al.,	)	
	)	
	)	
Defendants.	)	

BRIEF OF AMICUS CURIAE PROFESSOR JANE BAMBAUER IN SUPPORT OF PLAINTIFFS' COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

Christopher W. Weller
CAPELL & HOWARD, P.C.
150 South Perry Street
Montgomery, AL 3104
Phone: (334) 241-8066
Fax: (334) 241-8266
chris.weller@chlaw.com

Counsel for Amicus Curiae Professor Jane Bambauer

TABLE OF CONTENTS

TABLE OF AUTHORITIES

CASES

OTHER AUTHORITIES

REGULATIONS

STATUTES

INTEREST OF *AMICUS CURIAE*

Amicus is a Professor of Law at the University of Arizona and an expert in the public policy and industry practices related to privacy, research, and Big Data. Throughout my academic career, I have studied the societal risks and benefits related to the collection and use of personal data. Much of my scholarly and community service work relates to deidentified research data. In collaboration with statistical disclosure experts, I have written guidance documents, scholarly publications, and an amici curiae brief for the U.S. Supreme Court. I have worked with the ACLU of Arizona to facilitate public access to deidentified data on Border Patrol detainees. I have served on the Program Committee for UNESCO's annual conference on Privacy in Statistical Databases, and I have given presentations about the trade-off between privacy risk and research to the U.N. Economic Commission for Europe/Eurostat, the Federal Trade Commission, and Google.

I have no personal interest in the outcome of this case, but a professional interest concerning the impact that the adoption of Differential Privacy could have on government accountability and open research. As the government and private companies have access to increasing amounts of personally identifiable information, it is more important than ever that researchers, nonprofits, and journalists have access to accurate statistical data.

SUMMARY OF ARGUMENT

The State of Alabama has done an excellent job illustrating how the Census Bureau's use of Differential Privacy will affect the accuracy and reliability of nearly every statistical table and data product that is in use for highly consequential redistricting and resource allocation decisions. This amicus brief contributes a more fundamental critique and objection to Differential Privacy as a tool for mitigating risk in public datasets.

The Census Bureau's adoption of Differential Privacy is indefensible because the definition and measure of "privacy" imbedded in Differential Privacy is poorly matched to actual risk of disclosure. Because "privacy" is defined in a manner that is insensitive to context, including which types data are most vulnerable to attack, Differential Privacy compels data producers to make bad and unnecessary tradeoffs between utility and privacy. Reidentification attacks that are much more feasible, and thus much more likely to occur, are treated exactly the same as absurdly unlikely attacks. As a result, whatever "privacy" budget is chosen, the resulting noise-added data is simultaneously less accurate *and less privacy-protective* than a traditional disclosure control method that is attuned to context.

Thus, there is no rational basis for employing Differential Privacy. Differential Privacy, if used as intended, would wreak havoc on the accuracy of almost all US Census data products and defeat the very purpose for comprehensive Census data collection without any meaningful gain in the (already adequate) privacy protections. And it is particularly irrational given that the delays caused by implementing Differential Privacy will have serious consequences for elections this year. For these reasons, the adoption of Differential Privacy is an arbitrary and capricious use of the agency's discretion to balance competing societal interests in statistical accuracy and data privacy.

ARGUMENT

I. Differential Privacy Uses a Flawed Conception of Privacy

Differential Privacy guarantees to each data subject that the probability a statistical report will present a particular value is not too different from the probability that it would give the same value even if the data subject wasn't included in the dataset. As a practical matter, the guarantee requires a certain amount of noise (*i.e.*, the intentional introduction of precisely calibrated error)

to be added, and the amount of noise is determined by a worst-case scenario in which an attacker might know everything about a database except one last detail.

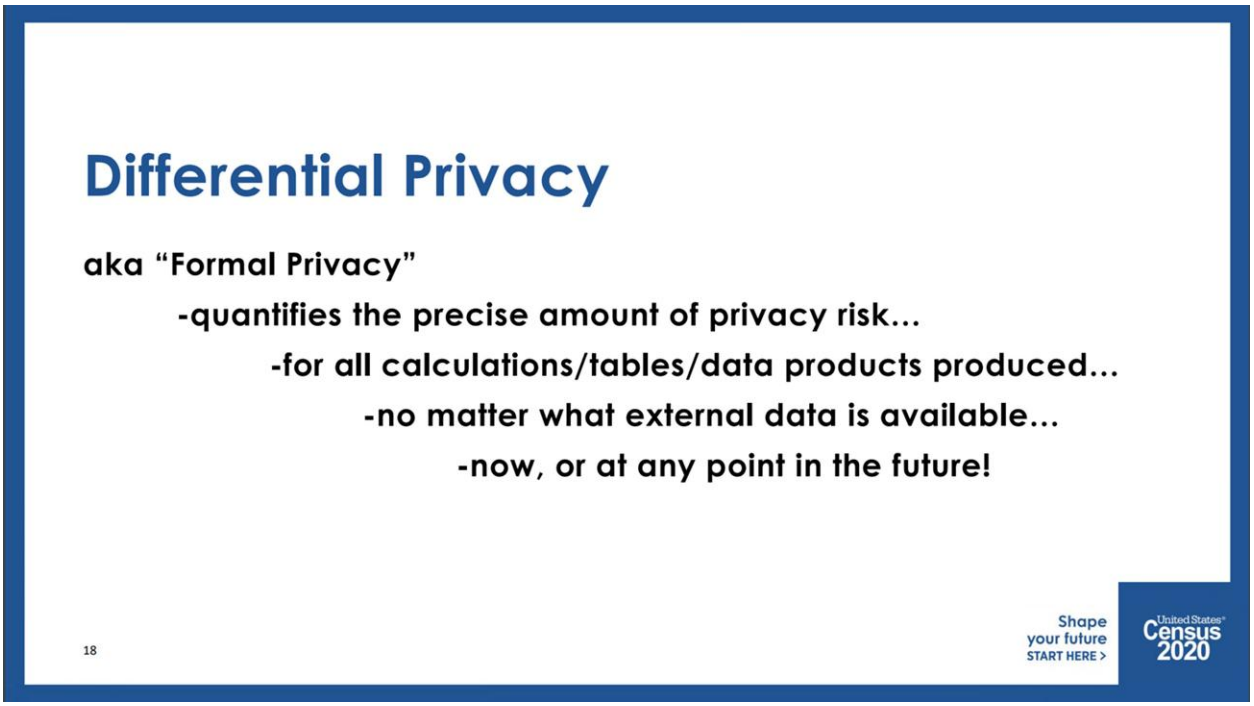
The Census Bureau chose to adopt Differential Privacy rather than continuing to use traditional disclosure control methods for two key reasons: Differential Privacy makes no assumptions about the reidentification attacks that could be possible now or in the future; and it quantifies the concept of "privacy" in a way that allows the Bureau to make and meet certain guarantees. However, each of these purported advantages of Differential Privacy is in fact detrimental to the Census Bureau's mission.

A. Differential Privacy Has No Relation to Real World Risk

Because Differential Privacy measures privacy under worst case scenarios, the privacy protections that are guaranteed by Differential Privacy are not dependent on context. No data steward has to make predictions or value judgments about which types of data are more vulnerable to reidentification attack, and which types are more sensitive and harmful if discovered. As the Census Bureau itself explains, Differential Privacy "does not directly measure re-identification risk (which requires specification of an attacker model). Instead, it defines the maximum privacy "leakage" of each release of information compared to some counterfactual benchmark (*e.g.*, compared to a world in which a respondent does not participate, or provides incorrect information.)"[1]

[1] Philip Leclerc, *The 2020 Decennial Census TopDown Disclosure Limitation Algorithm*, U.S. CENSUS BUREAU (Dec. 11, 2019) presentation available at https://www.nationalacademies.org/event/12-11-2019/docs/DCC854281ACE97996C107A2DC1BE711DFF02965EE0EC (last accessed Apr. 6, 2021).

This is characterized as a benefit by the Census Bureau as well as computer scientists who developed Differential Privacy since it automatically guards against every conceivable or hypothetical attack.[2]



However, the indifference to context is actually a drawback if the goal is to mitigate real world risk. The differential privacy model treats all data leakage the same, and all possible attacks as equally plausible. This is because privacy loss is measured based on an intruder who knows *everything* about *every person* except for one last piece information about one person.[3] Because the context-free definition of privacy leakage is so easily triggered, the privacy "guarantees" offered by Differential Privacy are deceptive. After all, in order to produce any useful data, the data steward must allow for *some* potential information leakage. The data steward does this by

[2] Michael Hawes, *Differential Privacy and the 2020 Decennial Census,* U.S. CENSUS BUREAU (Jan. 28, 2020) presentation available at https://zenodo.org/record/4122103/files/Privacy_webinar_1-28-2020.pdf (last accessed Apr. 6, 2021) (hereinafter "Hawes presentation").
[3] Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 54 COMM'NS OF THE ACM 89, 92 (2011).

selecting parameters like ε ("epsilon"), which allows some statistical data to be produced as long as the reports put a limit on the confidence that a nearly-omniscient attacker would have when receiving new information.[4] But these parameters do not and cannot ensure that the relaxations in privacy are well-aligned with real world risk. That is, if ε is large so as to allow reasonable levels of accuracy, it is just as likely to be "spent" on statistical products that we *know* are vulnerable to reidentification attack as it is on products that we have good reason to believe is not likely to be reidentified.

For example, when constructing the limited types of data that are available in enumeration district files, Differential Privacy requires the Census Bureau to protect against privacy leakage pertaining to Hispanic status. What this means is that noise must be added to thwart a hypothetical intruder who has access to the race, age, Census block, and housing type of a particular target as well as the race, age, Census block, housing type, and Hispanic status of *every single other person* in the target's district because this hypothetical intruder might then use the Census file to determine the Hispanic status of the target. The Census Bureau can of course spend some of its privacy budget to allow for more accurate reporting of data, but this privacy budget expenditure is wasteful. Nobody now or in the future will have access to that much auxiliary information in a form that reports exactly the same values as the Census data, and if they did, it's hard to believe they wouldn't know the Hispanic status of that last person. Yet by spending any part of a privacy budget to guard against this figment of the imagination, some other data table of high consequence will have to be made less accurate. If traditional disclosure control techniques can be criticized for

[4] For an illustrated explanation of Differential Privacy and the meaning of epsilon, *see* Jane Bambauer et al., *Fool's Gold: An Illustrated Critique of Differential Privacy,* 16 VAND. J. ENT. & TECH. 727 (2014).

failing to anticipate some types of attacks, Differential Privacy can be criticized for anticipating *all* of them.

A formal assumption that attackers will be virtually omniscient makes privacy protection easier for the Census Bureau because it relieves the agency from having to make educated (but uncertain) predictions about which types of threats are plausible and which are not. The adoption of Differential Privacy therefore shields the Census Bureau from criticism that the agency made errors in judging which threats were more or less plausible. But the same formalism that is held up as a benefit of Differential Privacy permits an abdication of the responsibility to assess risks realistically, and to use mitigating strategies (like the addition of noise) where they are most needed.[5]

Consider, for example, what would have happened if the Department of Health and Human Services had decided to implement Differential Privacy when it produced public data on COVID cases and hospitalizations. Even if data tables were produced only one a week (instead of daily) in

[5] Josep Domingo-Ferrer & Krishnamurty Muralidhar, *New Directions in Anonymization: Permutation Paradigm, Verifiability by Subjects and Intruders, Transparency to Users,* 337 INFO. SCIS. 11, 12-13, 18 (2016); Tapan K. Nayak et al., *Measuring Identification Risk in Microdata Release and Its Control by Post-Randomization,* CENTER FOR DISCLOSURE AVOIDANCE RESEARCH, U.S. CENSUS BUREAU (assessing the problem with formal privacy measures, like "differential privacy," and concluding "[t]hus, for developing practical disclosure control goals, it is essential for the agency to consider intruders with limited prior information about their target units."); Mark Elliot & Josep Domingo-Ferrer, *The future of statistical disclosure control*, 3.1, NAT'L STATISTICIAN'S QUALITY REV. INTO PRIVACY & DATA CONFIDENTIALITY METHODS (2018) ("Many authors have commented that this environment is inherently difficult—if not impossible—to understand and therefore directly assessing risk is itself impossible. This in turn has led to bad decision-making about data sharing (a strange mixture of over-caution and imprudence which is driven more often than not by the personality of the decision-maker rather than by rational processes.)"); Garret Christensen & Edward Miguel, *Transparency, Reproducibility, and the Credibility of Economics Research*, 56 J. OF ECON. LITERATURE 920, 969 (2018) ("They have established that there is inherently a trade-off between these two objectives (Dwork and Smith 2010; Heffetz and Ligett 2014), though few actionable approaches to squaring this circle are currently available to applied researchers, to our knowledge.").

order to preserve the "privacy budget," a year's worth of data on current hospitalizations and weekly case numbers would cause the data to be useless. Here, for example, is what a table of case counts and hospitalizations might look like for a sample of Alabama counties if the tables were produced using an epsilon of one (assuming that the department produces weekly tables, and *does not* produce any other data.)

Example of Differential Privacy Applied to COVID Data (epsilon = 1)

County	**Case #s this week**		**14-Day Change**		**Currently Hospitalized**		**14-Day Change**	
	True	**With DP**	**True**	**With DP**	**True**	**With DP**	**True**	**With DP**
Jefferson ›	420	237	-35%	-68%	157	146	8%	-51%
Madison ›	196	626	-34%	Infinite	62	0	-19%	0%
Montgomery ›	175	260	-11%	2500%	47	54	0%	32%
Tuscaloosa ›	168	126	-39%	-75%	21	215	-16%	Infinite
Mobile ›	140	215	-61%	41%	16	0	-62%	-100%
Shelby ›	140	253	-37%	26%	158	136	6%	27%
Baldwin ›	84	452	-47%	Infinite	56	197	-35%	Infinite
Lee ›	70	183	-31%	151%	9	0	-25%	0%
Talladega ›	63	118	-23%	Infinite	150	101	8%	-20%
Elmore ›	63	94	-55%	-58%	52	1	-13%	Infinite
Lauderdale ›	56	58	-18%	Infinite	5	41	0%	105%
Cullman ›	56	51	-29%	-50%	5	0	67%	0%
St. Clair ›	49	83	-9%	-48%	159	208	6%	-60%
Calhoun ›	49	264	-25%	Infinite	16	49	-30%	Infinite
Autauga ›	49	0	0%	-100%	65	102	-6%	Infinite
Marshall ›	49	55	17%	-53%	64	331	-19%	Infinite
Limestone ›	49	60	9%	-15%	65	9	-17%	-80%
Houston ›	49	0	40%	-100%	26	0	-13%	-100%
Chilton ›	35	0	-3%	-100%	60	265	0%	15%
Blount ›	35	0	0%	0%	148	227	9%	11%
Tallapoosa ›	35	0	13%	0%	9	0	-25%	-100%
Walker ›	35	97	-33%	62%	154	278	10%	Infinite
Morgan ›	28	86	-45%	-5%	75	0	-19%	-100%
Colbert ›	28	168	-15%	Infinite	7	0	-36%	0%
Etowah ›	28	0	-64%	-100%	13	13	-19%	Infinite
Jackson ›	21	0	-63%	-100%	135	227	2%	-15%
Russell ›	21	36	-68%	Infinite	40	46	-33%	-88%
Marion ›	14	0	-56%	-100%	4	86	100%	Infinite
Dale ›	14	12	-30%	-91%	29	122	-19%	Infinite
Coffee ›	14	0	27%	-100%	0	0	-100%	0%

Data sourced by the New York Times from the U.S. Department of Health & Human Services and state and local public health departments.

Even with a very generous "privacy budget" of 16 (the largest the U.S. Census Bureau has analyzed from its study of 2010 decennial data), several counties would miss critical trends or

harbor false senses of security and threat. And again, these tables assume that *no other data* will be reported. If the results were broken down by age, gender, or race, the error would be much worse.

Example of Differential Privacy Applied to COVID Data (epsilon = 16)

County	**Case # this week**		**14-Day Change**		**Currently Hospitalized**		**14-Day Change**	
	True	With DP	True	With DP	True	With DP	True	With DP
Jefferson ›	420	417	-35%	-36%	157	180	8%	28%
Madison ›	196	200	-34%	-37%	62	68	-19%	-14%
Montgomery ›	175	174	-11%	-13%	47	51	0%	24%
Tuscaloosa ›	168	165	-39%	-38%	21	20	-16%	-17%
Mobile ›	140	151	-61%	-58%	16	0	-62%	-100%
Shelby ›	140	156	-37%	-29%	158	150	6%	0%
Baldwin ›	84	85	-47%	-49%	56	65	-35%	-20%
Lee ›	70	68	-31%	-31%	9	0	-25%	-100%
Talladega ›	63	66	-23%	-23%	150	157	8%	8%
Elmore ›	63	71	-55%	-46%	52	37	-13%	-41%
Lauderdale ›	56	56	-18%	-19%	5	2	0%	-75%
Cullman ›	56	59	-29%	-20%	5	7	67%	-22%
St. Clair ›	49	36	-9%	-12%	159	164	6%	11%
Calhoun ›	49	47	-25%	-33%	16	25	-30%	0%
Autauga ›	49	49	0%	32%	65	65	-6%	-6%
Marshall ›	49	65	17%	44%	64	60	-19%	-13%
Limestone ›	49	36	9%	3500%	65	69	-17%	-8%
Houston ›	49	50	40%	-11%	26	26	-13%	-16%
Chilton ›	35	24	-3%	100%	60	63	0%	80%
Blount ›	35	16	0%	-67%	148	147	9%	2%
Tallapoosa ›	35	36	13%	44%	9	10	-25%	-9%
Walker ›	35	50	-33%	-2%	154	158	10%	36%
Morgan ›	28	20	-45%	-64%	75	78	-19%	-10%
Colbert ›	28	20	-15%	-13%	7	0	-36%	-100%
Etowah ›	28	22	-64%	-69%	13	20	-19%	33%
Jackson ›	21	17	-63%	-74%	135	145	2%	31%
Russell ›	21	20	-68%	-62%	40	33	-33%	-50%
Marion ›	14	11	-56%	-69%	4	0	100%	0%
Dale ›	14	15	-30%	-12%	29	43	-19%	19%
Coffee ›	14	18	27%	29%	0	0	-100%	-100%

The reason so much noise must be added to these tables in order to satisfy differential privacy is because the tables must be robust from an attack by a person who knows *every single person's COVID status in a given county except one person's* (the target's). Traditional methods of disclosure control would not make this preposterous assumption. Instead, with this limited data

(county-level geographic units, and no demographic data included), very little noise would be added, and that noise would focus on less populous counties or counties with a small number of hospitals and testing facilities.

One can analogize the Differential Privacy standard for privacy guarantees to the standards that courts had to develop in negligence cases to see the problem. The Census Bureau had been using statistical disclosure techniques that are consistent with the Hand formula from *United States v. Carroll Towing Co.*, 159 F.2d 169 (2d Cir. 1947). Risk was estimated based on the probability (p) that a misfeasor would have the auxiliary information to launch a successful attack and the losses (L) that would result from the disclosure of sensitive information. Risk under traditional notions of reasonableness would account for remote risks as well as common ones—threats that are very unlikely to materialize as well as those that are more common. But all would be appropriately weighted to reflect the probability and harm.

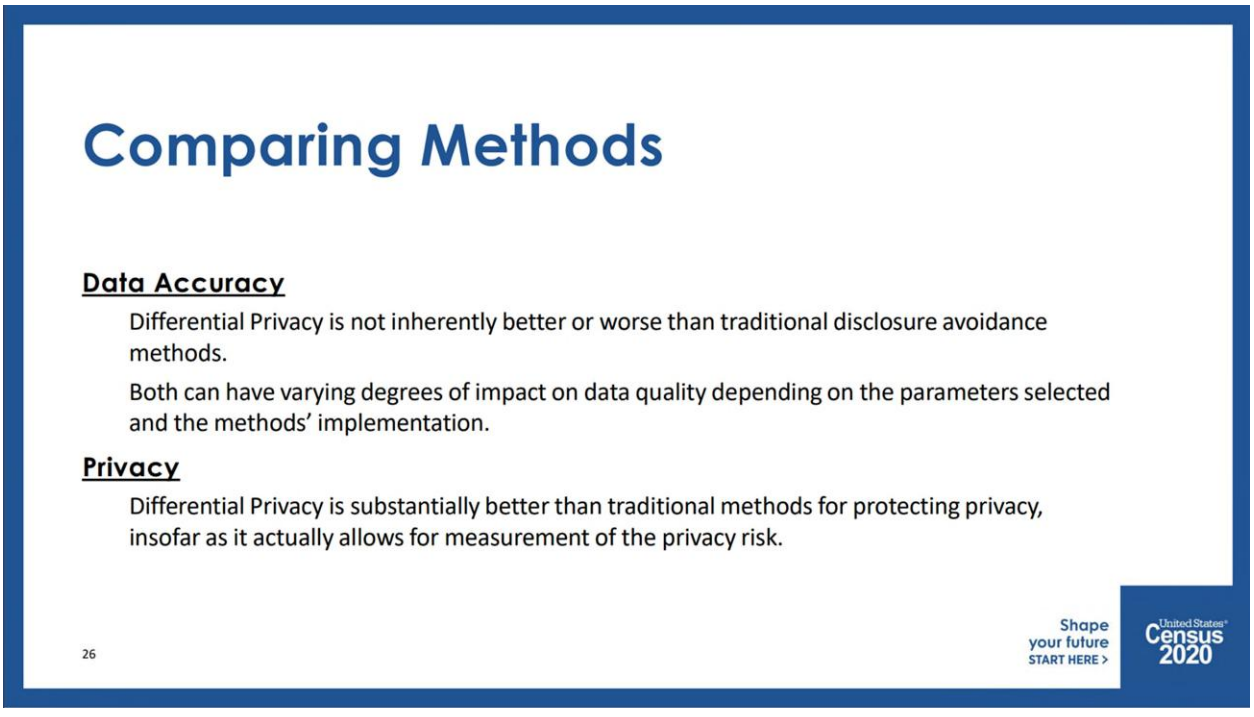
In contrast, by adopting Differential Privacy, the Census Bureau limits the public access and utility of Census data based on the worst-case hypotheticals. Differential Privacy guarantees are deliberately indifferent to real world considerations of risk. Differential Privacy defines privacy loss not based on what is foreseeable, but based on *the full universe* of hypotheticals. In the torts context, it would be equivalent to asking "if an omnipotent and all-powerful alien entered the scene, what could go wrong?" Indeed, the Census Bureau's own explanation of their definition of privacy risk assumes that an attacker "has infinite computing resources, infinitely powerful algorithms, and allows her to have arbitrary side knowledge."[6]

[6] Hawes presentation, *supra* note 2.

Arbitrary is the right word. The decision of the Census Bureau to abandon a standard of privacy protection based on foreseeable risks and to instead use a standard driven by nightmare fantasies is arbitrary and capricious, and an abuse of the Census Bureau's discretion.

B. Differential Privacy Provides a False Sense of Precision and Certainty

A second benefit of Differential Privacy, which is related to the first, is that it allows privacy to be measured without the error or uncertainty that comes with predicting which reidentification attacks are more or less feasible. It measures privacy loss in a theoretical sense, with mathematical certainty, rather than in an actuarial sense. The Census Bureau claims that Differential Privacy's ability to measure with certainty makes it "substantially better" than traditional methods for protecting privacy.[7]



However, the precision and certainty of Differential Privacy's measure of "privacy risk" is only valuable for measuring risk from theoretical worse case scenarios. In other words, the word

[7] Hawes presentation, *supra* note 2.

"privacy" in the Differential Privacy literature is a term of art that means something specific, and that does not account for probability. If the Census Bureau desired instead to measure risk with some attunement to the probability that an attack could be attempted or could succeed, Differential Privacy is inferior to the traditional methods that model different threat scenarios and quantify risks under a range of assumptions.

Thus, rather than being beneficial, the quantitative precision of Differential Privacy is actually a drawback. Differential Privacy has the patina of mathematical elegance without actually quantifying privacy risks of the sort that most people care about. Indeed, when I explain the meaning of privacy risk (or privacy loss) to lay audiences, people often respond that the privacy budget should depend on whether the variables disclosed in the statistical data are more vulnerable (large "p", in the Hand formula sense) or sensitive (large "L"). This, of course, is a reinvention of the disclosure avoidance techniques that the Census Bureau has used in the past and has now rejected with the adoption of Differential Privacy. Thus, it is useful to distinguish Differential Privacy's concept of abstract privacy loss from privacy *risks* based on probability and harm.

The precision of Differential Privacy's definition of "privacy" loss and its promise of privacy "guarantees" is also deceptive. Realistically, as the State of Alabama and its experts have shown, Census Bureau data cannot be produced in any useful form without using fairly generous values of the parameter ε. Indeed, when the Census Bureau produced an exemplary file of 2010 Census data with Differential Privacy techniques, the Bureau explained that it set a more "conservative" privacy-loss budget than it expects will be set for the 2020 census—meaning that the demonstration data had "more noise (error) than should be expected in the final 2020 Census

data products[.]"[8]This is welcome news for those who are concerned about accuracy, but the implicit result is that more accuracy will come at the cost of greater privacy loss. These losses may be trivial or they may be very risky under real world conditions. Differential Privacy does not distinguish between these.

II. Traditional Disclosure Control Techniques Do a Better Job Protecting Privacy and Preserving Utility

In the past, the U.S. Census Bureau successfully managed the risks inherent to public data releases using a range of disclosure control techniques. These methods often require data stewards to anticipate the most likely threats to data subjects, identify the most vulnerable records, and reduce the vulnerability with an eye toward preserving research potential. These techniques include data swapping, sampling, and blank-and-impute procedures that add uncertainty and error to the variables that are potentially vulnerable to reidentification attack.[9] Disclosure control is a highly pragmatic exercise that requires some grounded predictions of current and future behavior in order to make sure that the noise added to a dataset is strategically placed where a misfeasor is likely to attack. Privacy risk using these techniques is quantifiable, but requires some assumptions to be made about which attacks are remotely plausible and which are not.[10]

These techniques are not broken. Public use research datasets have continued to be safely produced without evidence of significant risk or harm to research subjects. Although there are

---

[8] 2020 Disclosure Avoidance System Updates, U.S. CENSUS BUREAU https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html (last accessed Apr. 6, 2021).

[9] Federal Committee on Statistical Methodology, Statistical Policy Working Paper 22 (2d Version, 2005).

[10] For a description of various methods to quantify privacy risk outside Differential Privacy, *see* David Sidi & Jane Bambauer, *Plausible Deniability,* 2020 PRIVACY IN STAT. DATABASES 91 (2020).

many studies and popular media reports that state deidentified data can be easily reidentified[11], the underlying research often relies on uniqueness of a data subject as the measure of reidentification risk, and simply assume attackers will possess ample information about their targets in identified form.[12] Moreover, even when an attacker *does* have significant amounts of auxiliary data, attacks are often so riddled with error that reidentifications are more likely to be wrong than right.

---

[11] Gina Kolata, *Your Data Were 'Anonymized'? These Scientists Can Still Identify You*, N.Y. TIMES (July 24, 2019), https://www.nytimes.com/2019/07/23/health/data-privacy-protection.html; Sophie Bushwick, *'Anonymous' Data Won't Protect Your Identity,* SCIENTIFIC AMERICAN (July 23, 2019), https://www.scientificamerican.com/article/anonymous-data-wont-protect-your-identity/; Kelsey Campbell-Dollaghan, *Sorry, Your Data Can Still Be Identified Even if It's Anonymized*, FAST COMPANY (Dec. 10, 2018), https://www.fastcompany.com/90278465/sorry-your-data-can-still-be-identified-even-its-anonymized; Rebecca Jacobson, *Your 'Anonymous' Credit Card Data Is Not So Anonymous, Study Finds*, PBS NEWS HOUR (Jan. 29, 2015, 5:54 PM), https://www.pbs.org/newshour/nation/anonymous-credit-card-data-anonymous-study-finds; Natasha Singer, *With a Few Bits of Data, Researchers Identify 'Anonymous' People,* N.Y. TIMES BITS (Jan. 29, 2015, 2:01 PM), https://bits.blogs.nytimes.com/2015/01/29/with-a-few-bits-of-data-researchers-identify-anonymous-people/.

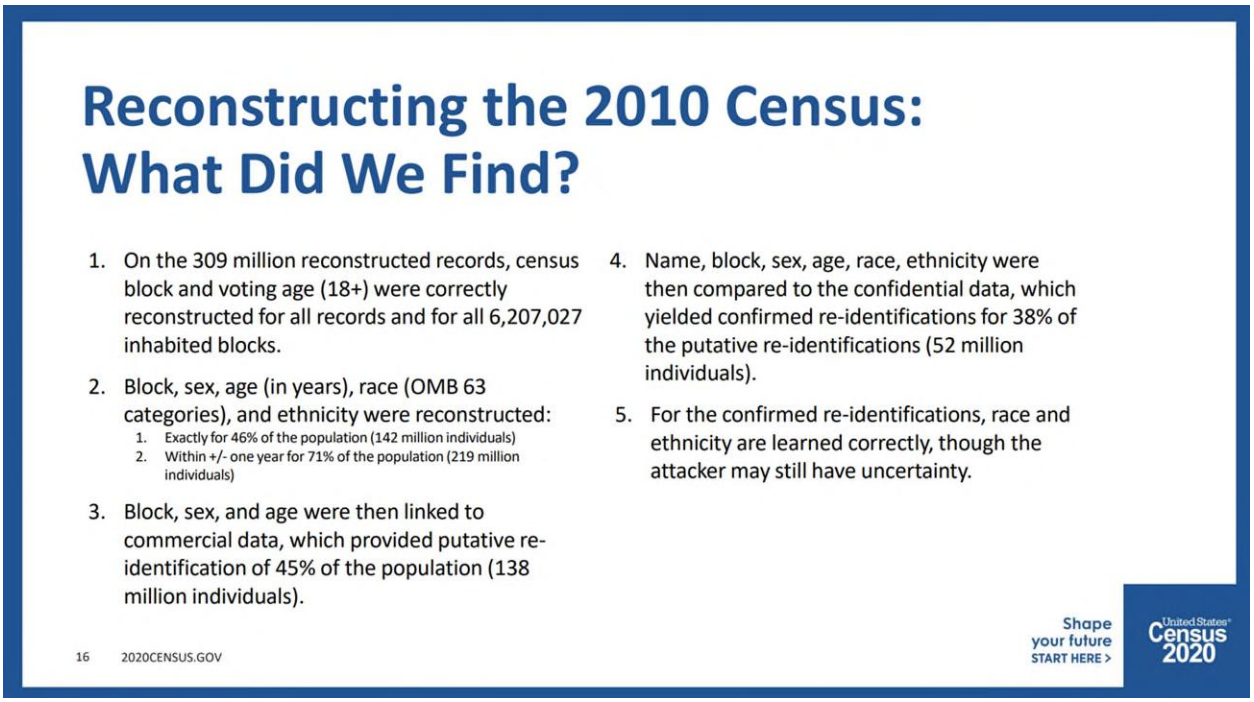
[12] Daniel Kondor et al., *Towards Matching User Mobility Traces in Large-Scale Dataset*, IEEE Transactions on Big Data (Vol. 6, Issue 4) (Dec. 1, 2020) (assessing "matchability" rather than reidentifiability, and finding an attacker could match 17% of the data subjects using "only" one week of comprehensive mobility information); Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference,* SCIENCE (Jan. 18, 2013) (concluding that intruders who already have the DNA sequence of a male relative might be able to identify a person in a genomic research database). For a critical take on using uniqueness as reidentification, *see* Gregory E. Simon et al., *Assessing and Minimizing Re-Identification Risk in Research Data Derived from Health Care Records,* 7 eGEMS 1, 3 (2019) ("To use a financial analogy, the exact amount (in dollars and cents) of the last 5 transactions in any credit account may be unique, but it would only be identifying to an adversary who already had access to those banking records."); Fida Kamal Dankar, *Estimating the Re-Identification Risk of Clinical Data Sets,* 12 BMC MED. INFORMATICS & DECISION MAKING 66 (2012); Luc Rocher et al., *Estimating the Success of Re-Identifications in Incomplete Datasets Using Generative Models,* 10 NATURE COMMS. art. 3069 (2019); Kathleen Benitez & Bradley Malin, *Evaluating re-identification risks with respect to the HIPAA privacy rule*, 17(2) J. AM. MED. INFOR. ASS'N 169 (2010) ("If a researcher receives a dataset drawn at random from the population of Ohio under Limited Dataset provisions, more than 1 out of 6 of those represented would be unique based on demographic information. Remember, though, that uniqueness is not sufficient to claim re-identification. There is still need for an identified dataset and *VOTER* reflects this reality. While higher than the risk under Safe Harbor, <*LIMITED, VOTER*> is significantly lower than <*LIMITED, GENERAL*>, particularly for smaller values of *g*. According to <*LIMITED, VOTER*>, only 0.002% of the population is 1-distinct and 0.01% is 5-distinct.")

Internal studies performed by the U.S. Census Bureau to test their past uses of disclosure control techniques demonstrate that traditional disclosure control techniques provide excellent protection against reidentification attacks. For example, in one study, a group of Census researchers attempted to attack the data from an individual-level public use dataset on over two million data subjects. The data subjects were selected from three counties that were specifically chosen because of their vulnerability (residents in these counties are less transient, and therefore less likely to have noisy or stale data.) Next, the researchers purchased identified data on 700,000 people in the selected counties from a data aggregator and used all available overlapping key variables such as age, ethnicity, gender, and income. Out of the more than 2 million records in the research data files, the researchers' matching algorithm made apparent matches on 389 individuals. However, of those 389 apparent matches, *only 87 were actually correct*—an accuracy rate of just 22%.[13] Most of the apparent matches were wrong.

The Census Bureau's more recent examination of the 2010 census records found greater numbers of apparent matches, but the attempted attacks were similarly lousy in making accurate matches. This time, the Census Bureau used all 309 million U.S. census records and used census block, sex, and age to match census records to a commercially available database. This time, the researchers were able to make matches on 45% of the records (a whopping 138 million individuals), presumably because of the value that block-level geographic area provides for making unique matches. However, the vast majority of those matches (62%) were *wrong*.[14]

[13] Ramachandran, *et al*., *Exploring Re-identification Risks in Public Domains*, U.S. CENSUS BUREAU (Sept. 12, 2012) accessible via https://www.census.gov/srd/papers/pdf/rrs2012-13.pdf.

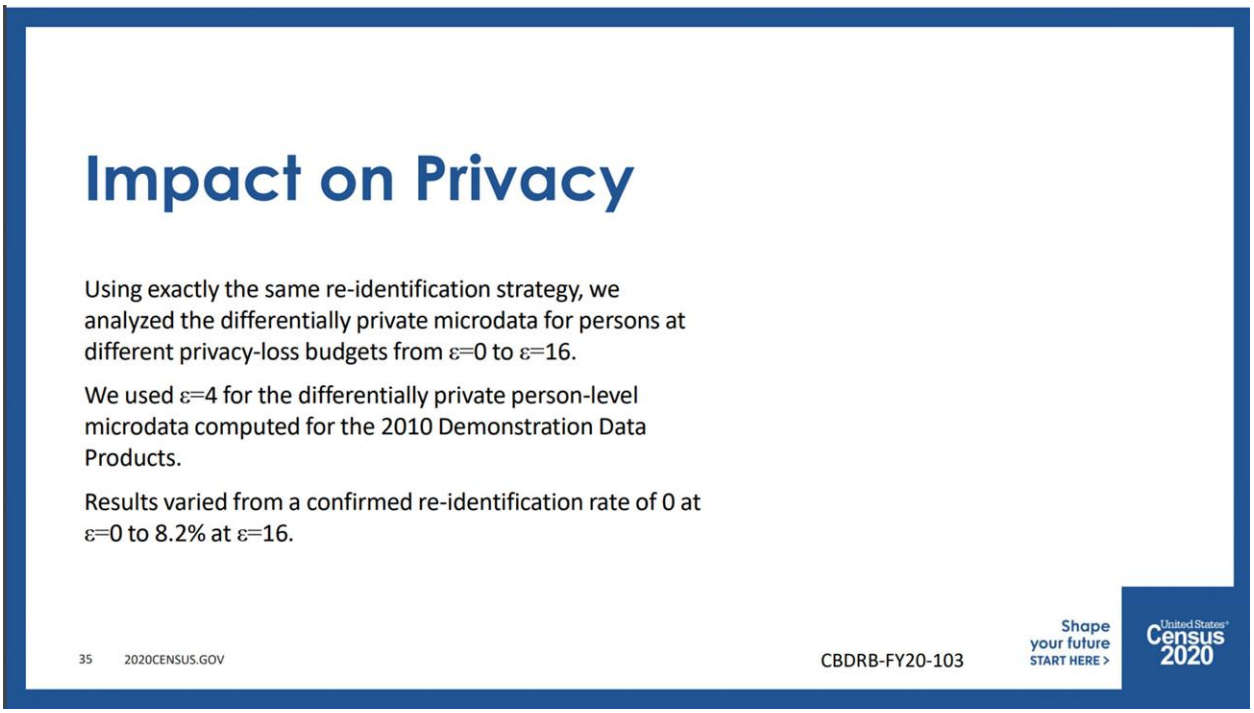
[14] Hawes presentation, *supra* note 2.



The Census Bureau presents this internal study as evidence that the Bureau needs to abandon traditional privacy methods and use Differential Privacy for the 2020 Census, but logic is strained. First, the conclusion that "the attacker may still have some uncertainty" is a dramatic understatement. It is misleading to suggest that 52 million individuals were accurately reidentified when the simulated attacker would not be able to distinguish them from the other 86 million individuals that the attacker falsely reidentified. The big numbers of reidentifications are meaningless if the Census Bureau credibly shows that even attacks that seem to succeed are most likely to be wrong.

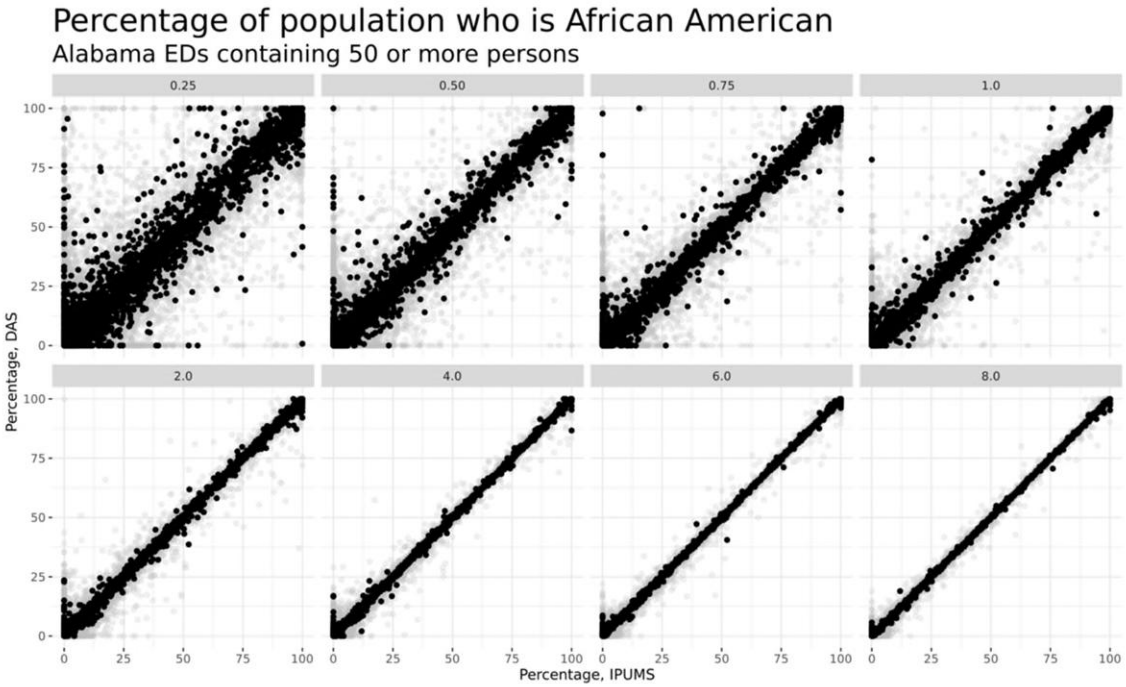
Moreover, when the Census Bureau applied the same simulation attack methods on data that it had prepared with Differential Privacy standards, confirmed reidentifications were in the same ballpark (about 25 million accurate reidentifications for an epsilon value of 16.) Thus, the advantages of switching to Differential Privacy are modest.[15]

[15] Hawes presentation, *supra* note 2.



By contrast, the disadvantages from Differential Privacy, in terms of utility loss, are severe. The State of Alabama and its expert witnesses have highlighted some of the difficulties that will arise from needless inaccuracies in the Census data, including the likelihood that flawed data will cause redistricting errors. Pls' Mot. for a Prelim. Injun., Doc. No. 3 at 35. The same flawed data will deprive nonprofit organizations of the opportunity to investigate or challenge voting rights violations, too. Given that the error in Black/African American residency can be off by hundreds in many of Alabama's key legislative districts (*Id.*), illegal redistricting would be hard to even allege, let alone prove. Even if the Census Bureau uses a larger "privacy budget" (epsilon of 6 or 8), we can still expect noise to cause minority representation to be over- or under-reported in a few districts, as David Van Riper, *et al*. demonstrated in their study of the 1940 differentially private decennial census files.[16]

[16] David Van Riper, et al., *Differential Privacy and the Decennial Census,* IPUMS DIFFERENTIAL PRIVACY WORKSHOP (Aug. 15, 2019) available at https://assets.ipums.org/_files/ipums/intro_to_differential_privacy_IPUMS_workshop.pdf.



These problems are bedeviling even for the relatively simple decennial census files, which do not report rich information about income and other sensitive, important traits. Once the Census Bureau begins to implement Differential Privacy to produce tables on income bands broken out by race, gender, and region, even a generous "privacy budget" will be spread so thin that the tables will become gibberish. Consider, again, the Alabama COVID tables presented above. If public health officials were to report the same figures broken down by age categories, race, and gender, the results would become even more erroneous. If the data is further segmented into smaller geographic or social units in order to understand whether, e.g., schools or nursing homes are having an outbreak, Differential Privacy would either prevent any meaningful statistics to emerge, or would require data stewards to select such a large "privacy budget" that realistic risks are unguarded.

The communities most likely to suffer from both the unjustified error and the unnecessary tolerance of (real world) privacy risk are small or vulnerable ones. Given that context-aware

assessments of privacy risk can outperform Differential Privacy, both for data utility and for protection from foreseeable threats, the Census Bureau's decision to use Differential Privacy is an unreasonable use of agency discretion.

By contrast, traditional disclosure control experts would add just enough noise to the table cells of counties with very small numbers of cases or with only a few sources of treatment and testing to cause error and uncertainty for the remotely plausible attacks in which a neighbor or doctor might already know nearly everybody who has tested positive for COVID. More noise or error would be introduced for tables that report on commonly known demographics or characteristics (such as race or status as a student) since an attacker could plausibly know the demographics and basic characteristics of the relevant population. But traditional disclosure control techniques would not have to anticipate that an attacker, say, knows the current and past COVID status of every individual in a county except one (or, possibly, even knows that last person's COVID status but does not know that target's race.) These attack scenarios, however, are treated as just as likely as any other. This is why so much noise must be added to tables of simple counts in order to meet the standards for Differential Privacy. And this is also why the decision to abandon disclosure control based on realistic threat models in favor of Differential Privacy is irrational.[17]

[17] Indeed, the Census Bureau's decision may not satisfy even constitutional rational basis review. Courts in recent years have found that regulations on the sale of caskets or on the practices of hair braiding studios had so little connection to societal welfare or risk mitigation that even these types of economic regulations imposed by statute were unconstitutional. *See St. Joseph Abbey v. Castille*, 712 F.3d 215 (5th Cir. 2013) (striking down a Louisiana law that gave funeral homes exclusive rights to sell caskets that the state attempted to justify by "abstraction for hypothesized ends"); *Brantley v. Kuntz*, 98 F. Supp. 3d 884 (W.D. Tex. 2015) (finding that regulations requiring salons to have sinks and certain types of equipment were irrational as applied to African hair braiding studio).

III. Neither Law Nor Public Distrust Can Justify the Census Bureau's Decision to Adopt Differential Privacy

Finally, there are no provisions in the U.S. Census Act that require the Census Bureau to take the action it has, nor are there any crises in public trust that can justify a dramatic shift in privacy protocols.

A. Privacy Laws

Privacy laws have long been crafted to allow data to be shared broadly or publicly in statistical, deidentified format despite the inherent risks involved. The "reasonableness" standard is the approach embodied in federal privacy regulations and industry guidance documents, and it is particularly well-matched to public data. HIPAA, for example, applies only to personal health information "(i) that identifies the individual; or (ii) [w]ith respect to which there is a reasonable basis to believe that the information can be used to identify the individual." 45 C.F.R. §169.103. Subsequent guidance and regulations make clear that traditional disclosure avoidance techniques meet the standard as long as individuals cannot be re-identified under realistic assumptions of threat. DHS guidance documents on HIPAA compliance do not require or even recommend the use of Differential Privacy.[18] (If they had, management of the pandemic would have been particularly chaotic.)

The language in the Census Act is similar to HIPAA's. The relevant confidentiality provision reads:

[18] DEPT. HEALTH & HUMAN SERVS., GUIDANCE REGARDING METHODS FOR DE-IDENTIFICATION OF PROTECTED HEALTH INFORMATION IN ACCORDANCE WITH THE HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT (HIPAA) PRIVACY RULE (2012), available at https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard.

> § 9. Information as confidential; exception
>
> (a) Neither the Secretary, nor any other officer or employee . . . may . . .
>
> (2) make any publication whereby the data furnished by any particular establishment or individual under this title *can be identified* … .

13 U.S.C. § 9 (emphasis added). Although the act does not contain the phrase "there is a reasonable basis to believe…" the operative language is nearly identical to HIPAA. Both laws ask whether information can be used to identify individuals. The phrase "reasonable basis for belief" provides a mental state requirement (negligence), but even if the Census Act intends to impose a strict liability regime on Census Bureau officers and employees, the task of assessing which types of data can or cannot be identified is the same as the HIPAA context. Moreover, if this weren't the case—if the phrase were meant to prohibit *any* publications that have any hypothetical chance of causing identification, Differential Privacy with any budget above 0 would violate the Act just as surely as traditional disclosure control techniques do.

It unlikely, however, Congress intended to impose a strict liability rule in any case. An excessively cautious approach to privacy would permit government agencies to evade public accountability and would close off the social benefits of public access. These are the core purposes of the Census Bureau.

B. Public Trust

The Census Bureau's adoption of Differential Privacy could also make sense if a spate of successful reidentification attacks warranted a new approach to data privacy, but there is no such history, and there is no public outcry about the statistical data products routinely released by the Census Bureau. If anything, the use of Differential Privacy could spur public distrust and resentment by injecting doubt in the accuracy and reliability of data used to allocate resources and

define voting districts. The Bureau is legally obligated not only to protect the confidentiality of the Census records, but also to protect the vitality and accuracy of the information in its possession. Section 181 of the US Census Act requires the census to produce "current, comprehensive, and reliable data" for state, county, and local government purposes. 13 U.S.C. § 181.

The public has as much interest in the reasonable accuracy of statistical census data as in its reasonable privacy. Indeed, the Census Bureau promises respondents that "Federal funds, grants and support to states, counties and communities are based on population totals and breakdowns by sex, age, race and other factors. Your community benefits the most when the census counts everyone. When you respond to the census, you help your community gets its fair share of the more than $675 billion per year in federal funds spent on schools, hospitals, roads, public works and other vital programs."[19]

Differential Privacy undermines the Census Bureau's mission of collecting and providing reliable and credible information, and as the public becomes aware of the significant damage done to the accuracy of data, a crisis in trust is likely to emerge. For example, the Census Bureau imposes several constraints on their use of Differential Privacy so that negative numbers of people are not reported. But the combination of non-negativity and the state-level population invariants consistently leads to bias in the reporting of counts for small subgroups.[20] To reduce the bias, at least one Census Bureau advisor has suggested the Bureau should consider dropping the non-negativity constraint even though "it may be confusing to say that a town has a negative, fractional

[19] U.S. Census Bureau, *Why a Census?: How the Census Benefits Your Community,* https://www.census.gov/programs-surveys/decennial-census/2020-census/about/why.html

[20] Barber Expert Report, Doc. 3-5 at 13-14 (explaining that "[t]he combination of the non-negativity constraint and population invariants consistently leads to bias increasing counts of small subgroups and small geographic units and decreasing counts of larger subgroups and geographic units." (citation omitted))).

number of individuals with a particular combination of uncommon attributes".[21] Negative numbers of people in the official statistics is more than confusing, though. Political fights are already suffering from a dearth of shared facts. By using Differential Privacy, the Census Bureau is putting one of the few sources of ground truth at risk. When Americans see Census Bureau reports with 141,000 Alabama children living without parents, *see* Declaration of Thomas Bryan, Doc. 3-6 at 11, distrust and low response rates are likely to ensue.

IV. The Census Bureau's Position Sets a Trap for Public Records Laws

The U.S. Census Bureau's claim that Differential Privacy is the *only* defensible way to keep statistical data safe sets a terrible precedent for government transparency and accountability more generally. The effect on public records laws could be devastating. If government agencies are able to justify their decisions to withhold records because they are not "differentially private" (even if they can be deidentified quite well), the landscape of public records and government accountability would change for the worse. These changes will offend American democratic values regardless of political identity. While public universities might use privacy exemptions to avoid public controversy related to Affirmative Action, law enforcement agencies will use those same exemptions to avoid public controversy related to racially biased policing.

For example, for several years the New York Civil Liberties Union suspected that the New York Police Department (NYPD) was using stop & frisk procedures in racially discriminatory ways. Aggregated reports had already verified that the number of police stops and frisks were growing, but the organization was not able to provide convincing evidence of racial bias until 2011, when the group successfully sued the NYPD under New York's freedom of information law

[21] Michael B. Hawes, *U.S. Census Bureau, Implementing Differential Privacy: Seven Lessons From the 2020 United States Census*, HARV. DATA SCI. REV., Issue 2.2 (Apr. 30, 2020), https://perma.cc/DB66-9B5R.

to gain access to an individual-level database documenting the stop and frisk program. This data provided strong circumstantial evidence of intensive policing of minorities without reasonable suspicion and without any meaningful gains in safety, and it provided the impetus and basis for a civil rights challenge against the NYPD.[22] Yet the data contained in the NYPD database could have been used to reidentify a stopped individual using the location, timing, and demographic characteristics of the individuals who were stopped.[23]

Other public records litigation has required police departments to provide footage from body-worn cameras (with faces blurred) and has required U.S. Customs and Border Patrol to provide redacted documents about individuals stopped at internal checkpoint or by roving patrol in order to facilitate citizen and journalist investigations of potential abuses.[24] Public records disclosures of individual-level data has allowed journalists to find flaws in state sex abuse criminal cases[25] and evidence that children from low-income households were excluded from public gifted

---

[22] *Stop-And-Frisk 2011*, NEW YORK CIV. LIBERTIES UN. (May 2012), https://www.nyclu.org/sites/default/files/publications/NYCLU_2011_Stop-and-Frisk_Report.pdf; *Stop-and-Frisk in the de Blasio era*, NEW YORK CIV. LIBERTIES UN. (Mar. 2019) https://www.nyclu.org/sites/default/files/field_documents/20190314_nyclu_stopfrisk_singles.pdf ; *Floyd v. City of New York*, 959 F. Supp. 2d 540 (S.D. N.Y. 2013).

[23] Footage of several NYPD stops are available on YouTube, some with date and location information.

[24] Jessie Gomez, *Louisiana judge grants access to state policy body-camera footage*, MUCKROCK (Mar. 1, 2019) https://www.muckrock.com/news/archives/2019/mar/01/louisiana-bodycam/; James Lyall, et al., *Record of Abuse, Lawlessness and Impunity in Border Patrol's Interior Enforcement Operations*, AM. CIV. LIBERTIES UN. OF ARIZ., 4 (Oct. 2015); *ACLU Found. of Ariz. v. U.S. Dep't Homeland Sec.*, No. CV-14-02052-TUC-RM (BPV), 2017 WL 8895339 (D. AZ. Jan. 26, 2017) (rejecting the government's reidentification risk argument).

[25] Tennessee Watson, *Justice Isn't Always Done for Child Sex Abuse-I Know Firsthand*, REVEAL (Aug. 11, 2016), https://revealnews.org/article/tennessee-watson-justice-isnt-always-done-for-child-sexual-abuse-i-know-firsthand/.

and talented education programs.[26] Public access to data of this sort will be under grave threat if state agencies are able to say, as the Census Bureau has, that accuracy must be compromised for the sake of an abstract and baffling concept of privacy.[27]

CONCLUSION

All statistical data carries risk of inadvertent disclosure. Those who prepare public data must find a sensible way to balance the risks of privacy invasion against the risks of not allowing public access and accountability. The disclosure avoidance literature routinely acknowledges that data de-identification is a balancing act between data privacy and research utility, and it has served the U.S. Census very well up to this point.[28] Differential Privacy introduces unreasonable amounts

---

[26] Joseph Neff, Ann Doss Helms, & David Raynor, *Why Have Thousands of Smart, Low-Income NC Students Been Excluded from Advanced Classes?*, THE CHARLOTTE OBSERVER (May 21, 2017), https://www.charlotteobserver.com/news/local/education/article150488822.html.

[27] Government agencies have already used privacy as an excuse to withhold domestic violence data and now-infamous photographs from Abu Ghraib. Caitlin Russell, *In Massachusetts, laws intended to protect domestic abuse victims' privacy are being used to deny access to data about enforcement*, MUCKROCK (Jan. 9, 2018), https://www.muckrock.com/news/archives/2018/jan/09/dv-mass-data/; *ACLU v. Dep't of Defense*, 543 F.3d 59, 84 (2d Cir. 2008) ("According to the defendants, when combined with information contained in the investigative reports associated with the detainee images, release of the photographs could make it possible to identify the detainees.")

[28] "Stewards of social data [] face a fundamental tension. At one extreme, a data steward could share a complete dataset publicly with everyone. This *full release* approach maximizes the potential for scientific discovery, but it also maximizes risk to the people whose information is in the dataset. At the other extreme, a data steward could share the data with no one. This *no release* approach minimizes risk to participants, but it also eliminates benefits that could come from the responsible use of the data. In between these two extremes—no release and full release—there are a variety of intermediate solutions, which involve balancing risk to participants and benefits to science." Ian Lundberg, et al., *Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge* (Sept. 10, 2019), https://journals.sagepub.com/doi/10.1177/2378023118813023; *See also* Mark Elliot & Josep Domingo-Ferrer, *The Future of Statistical Disclosure Control,* NAT'L STATISTICIAN'S QUALITY REV. (2018) ("SDC fundamentally consists of two processes: disclosure risk analysis and disclosure control. Controlling the disclosure risk must be done in a way that optimizes the trade-off between risk and utility. While risk must be kept below the maximum acceptable threshold (set by law or by good practices), utility must be kept above the minimum threshold that data users can accept. Without utility constraints, there would be no reason to control disclosure: one might rather

of inaccuracy while making allowances for privacy loss that are not based on real world risk. Thus, the switch from grounded risk-based privacy precautions to the abstract guarantees provided by Differential Privacy is an arbitrary and capricious abuse of the Census Bureau's discretion. Although the Census Bureau has significant freedom to exercise its judgment over how best to balance privacy and data utility, the agency still "must examine the relevant data and articulate a satisfactory explanation for its action including a rational connection between the facts found and the choice made." *Motor Vehicle Mfrs. Ass'n of U.S., Inc. v. State Farm Mut. Auto. Ins. Co.*, 463 U.S. 29, 30 (1983). Because Differential Privacy is, by design, insensitive to real world probabilities and risks, and because there is no history of significant privacy breaches that would be corrected by a change to Differential Privacy, the Census Bureau cannot meet even the generous standard that applies to agency discretionary judgments.

Respectfully submitted,

/s/ Christopher W. Weller
CHRISTOPHER W. WELLER (WEL020)

Counsel for Amicus Curiae
Professor Jane Bambauer

OF COUNSEL:
CAPELL & HOWARD, P.C.
150 South Perry Street
Montgomery, AL 3104
Phone: (334) 241-8066
Fax: (334) 241-8266
chris.weller@chlaw.com

suppress the data entirely, which would result in 0% disclosure risk!"); El Emam & Luk Arbukle, ANONYMIZING HEALTH DATA: CASE STUDIES AND METHODS TO GET YOU STARTED 28 (2013) ("Zero risk can't guarantee if we want to share any useful data. The very small risk is the trade-off we need to accept to realize the many important benefits of sharing and using health data… Regulators don't expect zero risk either—they accept that a very small risk is reasonable.")

CERTIFICATE OF SERVICE

I hereby certify that on 9th day of April, 2021, I filed with the Court and served on all counsel through the CM/ECF system the foregoing document.

STEVE MARSHALL
Attorney General of Alabama
Edmund G. LaCour Jr. (ASB-9182-U81L)
Solicitor General
A. Barrett Bowdre (ASB-2087-K29V)
Deputy Solicitor General
James W. Davis (ASB-4063-I58J)
Winfield J. Sinclair (ASB-1750-S81W)
Brenton M. Smith (ASB-1656-X27G)
Assistant Attorneys General

STATE OF ALABAMA
OFFICE OF THE ATTORNEY GENERAL
501 Washington Ave.
Montgomery, AL 36130
Telephone: (334) 242-7300
Fax: (334) 353-8400
Edmund.LaCour@AlabamaAG.gov
Barrett.Bowdre@AlabamaAG.gov
Jim.Davis@AlabamaAG.gov
Winfield.Sinclair@AlabamaAG.gov
Brenton.Smith@AlabamaAG.gov
Counsel for the State of Alabama

Jason B. Torchinsky (VA Bar No. 47481)*
Jonathan P. Lienhard (VA Bar No. 41648)*
Shawn T. Sheehy (VA Bar No. 82630)*
Phillip M. Gordon (VA Bar. No. 95621)*
HOLTZMAN VOGEL JOSEFIAK
TORCHINSKY, PLLC
15405 John Marshall Hwy
Haymarket, VA 20169
(540) 341-8808 (Phone)
(540) 341-8809 (Fax)
Jtorchinsky@hvjt.law
Jlienhard@hvjt.law
Ssheehy@hvjt.law
Pgordon@hvjt.law
**pro hac vice*
Counsel for Plaintiffs

BRIAN M. BOYNTON
Acting Assistant Attorney General
ALEXANDER K. HAAS
Director, Federal Programs Branch
BRAD P. ROSENBERG
Assistant Director, Federal Programs Branch
ZACHARY A. AVALLONE
ELLIOTT M. DAVIS (N.Y. Reg. No. 4596755)
JOHN ROBINSON
Trial Attorneys
Civil Division, Federal Programs Branch
U.S. Department of Justice
1100 L St. NW
Washington, DC 20005
Phone: (202) 514-4336
Fax: (202) 616-8470
E-mail: elliott.m.davis@usdoj.gov
Counsel for Defendants

/s/ Christopher W. Weller
Of Counsel