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

THE STATE OF ALABAMA, *et al.*,	)	
	)	
Plaintiffs,	)	
	)	
v.	)	Civil Action No.
	)	3:21-CV-211-RAH
UNITED STATES DEPARTMENT OF COMMERCE, *et al.*,	)	
	)	
	)	
Defendants.	)	

AMICUS BRIEF OF DATA PRIVACY EXPERTS

Ryan Calo
Ran Canetti
Aloni Cohen
Cynthia Dwork
Roxana Geambasu
Somesh Jha
Nitin Kohli
Aleksandra Korolova
Jing Lei
Katrina Ligett
Deirdre K. Mulligan
Omer Reingold
Aaron Roth
Guy N. Rothblum
Aleksandra (Seša) Slavkovic
Adam Smith
Kunal Talwar
Salil Vadhan
Larry Wasserman
Daniel J. Weitzner

Shannon L. Holliday
(ASB-5440-Y77S)
COPELAND, FRANCO, SCREWS & GILL, P.A.
P.O. Box 347
Montgomery, AL 36101-0347

Michael B. Jones
Georgia Bar No. 721264
jones@bmelaw.com
BONDURANT MIXSON & ELMORE, LLP
1201 West Peachtree Street, NW
Suite 3900
Atlanta, GA 30309

Counsel for the Data Privacy Experts

TABLE OF CONTENTS

STATEMENT OF INTEREST

Amici are leading experts in data privacy and cryptography, and the connections of these fields to machine learning, statistics, and information theory. Amici's research and expertise with both differential privacy and the database reconstruction techniques used for reconstruction-abetted re-identification attacks offer a particularly well-informed perspective on the technical issues presented in this case.

SUMMARY OF ARGUMENT

Amici, listed in Appendix A, submit this brief to provide the Court with a fuller understanding of the risks of reconstruction-abetted re-identification attacks, and the unique role that differential privacy plays in protecting statistical releases against them. This case is about the capacity of the Census Bureau to honor its confidentiality commitment in light of new and evolving threats. We offer the Court additional information about the prevalence of reconstruction attacks, the growing ease with which they can be undertaken, and the risks they pose to the privacy of census participants and therefore to the census and the important public purpose it serves.

Together, Amici have developed reconstruction attacks, proved that they are a mathematical certainty, and co-invented the only known methodology for addressing them. We write to assure this Court that the Census Bureau's decision to use differential privacy is sound and essential and reflects the widely shared understanding across the field that it is the only method available to protect statistical releases from reconstruction attacks. We also seek to clarify two technical points. *First*, differential privacy and the 2020 Disclosure Avoidance System (2020 DAS) are distinguishable. *Second*, based on available information about the 2020 DAS, it does not use statistical inference.

ARGUMENT

I. Reconstruction Attacks Are Real and Put the Confidentiality of Individuals Whose Data Are Reflected in Statistical Disclosures at Serious Risk.

To appreciate the critical need for the use of differential privacy in the protection of census data, it is vital for the Court to understand the threat posed by reconstruction attacks and the re-identification attacks they facilitate.

As the Census Bureau's research—as well as extensive academic research—shows, reconstruction and the more familiar re-identification attacks can go hand in hand: first, attackers reconstruct person-level data records from products based on aggregated personal data, then, re-identify the reconstructed records.[1] Data releases protected by traditional statistical disclosure limitation techniques are vulnerable to these attacks. Thus, traditional statistical disclosure limitation techniques are no longer adequate to meet the Census Bureau's obligation to maintain the confidentiality of individual census responses.

Although the data produced by reconstruction and re-identification attacks contain some misidentified records and uncertainty, the data still poses a threat. Many records resulting from such attacks are accurately identified. In addition, the risks of reidentification attacks are not evenly distributed throughout the population. Individuals with less common attributes or combinations of them are at greater risk of exposure.

[1] *See infra* Section I.

A. Overview of Reconstruction Attacks.

Reconstruction attacks are processes for deducing highly accurate approximations of individual-level data from aggregated statistics.[2] There is a rich mathematical literature showing that reconstruction attacks pose a particular threat when the aggregated statistics consist of numerous simple counts like those published by the Census (e.g., the number of people in each census block broken down by race, ethnicity, and voting age).[3] The simple mathematical fact is that privacy loss from aggregate statistics works like radiation exposure: small, individually innocuous dosages of privacy erosion from published statistics accumulate until large-scale reconstruction is possible.

Reconstruction attacks reverse disclosure avoidance. Using only the information published in statistical reports, an attacker is able to deduce large swaths of the underlying confidential person-level data records with high accuracy.[4] These reconstructed records can then be identified—tied to an individual—by linking to commercial datasets using standard techniques. Therefore, if a purported disclosure avoidance technique allows reconstruction, then it does not meaningfully avoid disclosure.

Although reconstruction attacks are a new threat, carrying them out no longer requires significant innovation on the part of the attacker. The mathematical framework for reconstruction

[2] Cynthia Dwork, Adam Smith, Thomas Steinke & Jonathan Ullman, *Exposed! A Survey of Attacks on Private Data*, Annu. Rev. Stat. Appl. 4:12.1, 12.4 (2017).

[3] *Id.* at 12.4-12.6.

[4] *See generally id.* at 12.5-12.6.

attacks was discovered in 2003.[5] The technical community's understanding of such attacks was strengthened and generalized by subsequent work.[6]

B. The Census Bureau's Reconstruction Attack Demonstration.

Aware of the developing literature on reconstruction attacks, the Census Bureau appropriately sought to gauge how "at risk" census data was to such attacks. As described in the declaration of Chief Scientist and Associate Director for Research and Methodology John M. Abowd, researchers from the Census Bureau revealed the results of a reconstruction attack on the 2010 Disclosure Avoidance System.[7] An internal research team was able to completely reconstruct much of the confidential data underlying statistical publications from the 2010 census.[8] Using only a small fraction of the summary statistical tables released to the public, the researchers were able to reconstruct the underlying database of "person-level" data with a high degree of accuracy, revealing records for all 308,745,538 individuals counted in the 2010 Census, including the

[5] Irit Dinur & Kobbi Nissim, *Revealing Information While Preserving Privacy*, Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 1, 202-10 (2003), http://doi.acm.org/10.1145/773153.773173.

[6] *See generally* Aloni Cohen & Kobbi Nissim, *Linear Program Reconstruction in Practice*, 10 J. of Privacy and Confidentiality 1 (2020), https://doi.org/10.29012/jpc.711; Cynthia Dwork, Frank McSherry, and Kunal Talwar, *The Price of Privacy and the Limits of LP Decoding*," STOC '07: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing, June 2007, at 85–94, lpdecoding.pdf (kunaltalwar.org); Shiva Kasiviswanathan, Mark Rudelson, Adam Smith & Jonathan Ullman, *The Price of Privately Releasing Contingency Tables and the Spectra of Random Matrices with Correlated Rows*, STOC '10: Proceedings of the Forty-Second ACM Symposium on Theory of Computing, June 2010, at 775-784, https://doi.org/10.1145/1806689.1806795; Cynthia Dwork and Sergey Yekhanin, *New Efficient Attacks on Statistical Disclosure Control Mechanisms*, Advances in Cryptology – CRYPTO 2008, 1, 469-80 (David Wagner ed., 2008), https://doi.org/10.1007/978-3-540-85174-5_26.

[7] Decl. of John M. Abowd ¶ 38, Doc. 41-1.

[8] *Id.*

individual's "[b]lock, sex, age, race, [and] ethnicity."[9] The attack accurately reconstructed these fields for approximately 219 million people, or 71% of the population (to within one year of age), with exact reconstruction on 46% of the population, or 142 million people.[10] The Census Bureau's reconstruction demonstration closely followed the framework described by Dinur and Nissim. As described *infra*, the Census Bureau used the reconstructed data and commercially available datasets to successfully exactly reconstruct and re-identify the confidential census responses of 52 million people, without using any confidential Census data.

Using publicly and commercially available data and using eighteen-year-old techniques, the 2010 decennial census responses of 52 million people were reconstructed and re-identified. This is more than the combined 2010 enumerated population of the States of Alabama, Texas, and Florida—the Plaintiff's and the two largest amici states, respectively. If they are as vulnerable as the average census respondent, reconstruction and re-identification would expose the private Census responses of 91 members of the United States Congress, 23 members of the Alabama Legislature, and 24 sitting federal district and appellate judges within the Eleventh Circuit.

Plaintiff, Amicus Curae Jane Bambauer, and expert witness Steve Ruggles downplay the seriousness of this demonstration. The latter contrasts the Census Bureau's reconstruction results with a "simple simulation" of what can be predicted through chance alone.[11] Ruggles asks: What fraction of the population's 2010 census responses could be randomly guessed, rather than reconstructed? But Ruggles' analysis compares only on coarse statistics and vastly understates the real effectiveness of the Census attack.

---

[9] *Id.* at ¶ 38 & 108.

[10] *Id.*

[11] Ruggle's Expert Report, Appendix A, Page 7.

Ruggles does not separate out the rate of reconstruction according to Census block size.[12] The Census Bureau reconstruction does surprisingly well even on blocks of size 0-9: 20+% success; it achieves over 40% exact matches on blocks of size 10-49.[13] More than 32% of Alabama residents live in blocks of size 10-49.[14] On these blocks, Ruggles' random guessing has an inferior success rate of 12-15%. On blocks of size 0-9 its success rate is 3.5%. These comparisons are generous to Ruggles' random guessing; for example, census reconstructs age, sex, race, ethnicity, and block. Ruggles' guessing algorithm is given the block and guesses only age and sex.

Differential Privacy says that the risk of any harm remains essentially unchanged, independent of whether one joins or refrains from joining a dataset."[15] For residents of large blocks, participation in the Census will not substantially affect their likelihood of being reconstructed via random guessing. But without Differential Privacy, residents of small blocks will indeed suffer increased risk of reconstruction by participating in the Census. The Census Bureau has an obligation to protect the more than 1.7 million Alabamans living in small blocks.

---

[12] The size of a block matters. It's easier to randomly guess a card in your opponent's hand when playing Thirteen Card Rummy than when playing Three Card Poker. In the same way, random guessing works very well in blocks with hundreds or thousands of people, but very poorly for blocks with just tens of people.

[13] Appendix B of Declaration of John Abowd, Figure 1.

[14] *Id.*

[15] "The Mete and Measure of Privacy", Lecture by Cynthia Dwork, 152nd Annual Meeting of the National Academy of Sciences, April 2015, Research Briefings: April 25, 2015 (nasonline.org). The mathematical consequence is that algorithms operating on datasets should behave similarly on datasets that differ in the data of a single individual. How the *algorithm behaves* has nothing to do with what a privacy *adversary knows*, so Differential Privacy automatically protects against arbitrarily knowledgeable adversaries. This is the worst-case protection that Bambauer derides.

These small blocks are exactly those where attacks are most problematic, and where the protections of differentially private methods are most meaningful.

Reconstruction (and subsequent re-identification) of Census records does not require access to confidential Census records nor the expertise and computational resources of a federal agency. Columbia University Professor of Journalism Mark Hansen, working with a graduate student in statistics, "were able to perform our own reconstruction experiment on Manhattan. Roughly 1.6 million people are divided among 3,950 census blocks—which typically correspond to actual city blocks. The summary tables we needed came from the census website; we used simple tools like R and the Gurobi Optimizer; and within a week we had our first results."[16] This attack used an academic version of Gurobi;[17] a commercial version would be much faster.

C. Other Reconstruction Attack Demonstrations.

The same approach was used in 2018 to power another reconstruction attack, this one against a commercial disclosure avoidance system called Diffix.[18] Diffix was advertised as a system that provides off-the-shelf compliance with Europe's General Data Protection Regulation (GDPR).[19] According to its creators, "the French national data protection authority" had evaluated Diffix and found that it "delivers GDPR-level anonymity."[20] Cohen and Nissim adapted the 2003 reconstruction attack blueprint to Diffix and, with a few hundred lines of code running in less than

[16] Mark Hansen, *To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data*, N.Y. Times (Dec. 5, 2018), https://www.nytimes.com/2018/12/05/upshot/to-reduce-privacy-risks-the-census-plans-to-report-less-accurate-data.html.

[17] *Id.*

[18] Cohen & Nissim, *supra* note 7 at 3-4.

[19] *Id.*

[20] *Id.*

ten seconds on a laptop, perfectly reconstructed the data without any error.[21] What happened next is typical of the patch-break-patch again cycle that plagues traditional approaches to disclosure limitation that do not have rigorous guarantees: the company behind Diffix updated their system and claimed to defend against the attack.[22] However, it was quickly shown that a slight modification of the same attack could still perfectly reconstruct the social security numbers of about 90% of data subjects.[23]

A reconstruction attack on statistical reports released by the Israel Central Bureau of Statistics (CBS) was carried out in 2014.[24] CBS conducts an annual Social Survey, with questions about religion, ethnicity, employment, education, income, family, health, and attitudes, among many others.[25] CBS made these statistical reports publicly available online.[26] Undergraduate computer science students demonstrated that they could completely reconstruct the survey responses of over 14% of data subjects—1005 out of the 7064 survey subjects.[27] The students stopped reconstructing the data after they re-identified one of the survey subjects—an acquaintance of one of the students.[28]

---

[21] *Id.*

[22] *Id.*

[23] Aloni Cohen, Sasho Nikolov, Zachary Schutzman & Jonathan Ullman, *Reconstruction Attacks in Practice*, DifferentialPrivacy.org (Oct. 27, 2020), https://differentialprivacy.org/diffix-attack/.

[24] Amitai Ziv, *Israel's 'Anonymous' Statistics Surveys Aren't So Anonymous*, Haaretz (Jan. 7, 2013), https://www.haaretz.com/surveys-not-as-anonymous-as-respondents-think-1.5288950.

[25] *Id.*

[26] *Id.*

[27] *Id.*

[28] *Id.*

These examples are not flukes, but evidence of a new and growing risk facing statistical agencies. The ability to reconstruct data records from overly accurate statistics is a mathematical certainty. A consensus study report published by the National Academies of Sciences, Engineering, and Medicine in 2017 concluded that traditional statistical disclosure methods "are increasingly susceptible to privacy breaches given the proliferation of external data sources and the availability of high-powered computing that could enable inferences about people or entities in a dataset, re-identification of specific people or entities, and even reconstruction of the original data."[29] The research described above has born this out.

D. Reconstruction Attacks Enable Re-Identification Attacks.

Reconstruction yields accurate records for a large swath of Census respondents, with names removed. Such data are often called "anonymized." But "re-identification" of anonymized data is notoriously common and will be easier when leveraged by more modern datasets. Re-identification is the process of associating person-level data to the identities of actual people.[30] Once records are reconstructed, re-identification is relatively easy. It uses standard and well-known techniques and requires only access to commercial or public datasets that overlap with the anonymous records on some subset of the data fields. As the President's Council of Advisors on Science and Technology wrote in 2014, "it is increasingly easy to defeat anonymization by the very techniques that are being developed for many legitimate applications of big data."[31]

---

[29] National Academies of Sciences, Engineering, and Medicine, Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps 104-05 (2017).

[30] Boris Lubarsky, *Re-identification of "Anonymized Data,"* 1 Geo. L. Tech. Rev. 202, 208-09 (2017).

[31] President's Council of Advisors on Science and Technology, *Report to the President, Big Data and Privacy: A Technology Perspective* (2014),

Examples of re-identification of anonymized health records abound. Sweeney re-identified patients in anonymized health records from Washington state using newspaper stories[32]; Yoo et. al. re-identified patients in Maine and Vermont using newspaper stories as well, even when such data is anonymized according to the principles set forth in the HIPAA Safe Harbor Standard[33]; and Sweeney et. al. re-identified individuals from the Northern California Household Exposure Study using the combination of tax data and online tools—such as a data broker website and Google Earth and Street View—even when the data were anonymized to standards beyond what is required by HIPAA's Safe Harbor Standard.[34] In August 2016, the Australian Government released three billion records of billing data from its Medicare and Pharmaceutical Benefits Schemes, covering 10% of the Australian population (2.9 million people), anonymizing not only the patient but the medical provider as well.[35] Shortly after release, researchers re-identified all medical providers in the dataset.[36] Separately and without using re-identified provider information, the researchers manually re-identified at least five patients by linking approximate birth dates of children in the

---

https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.

[32] Latanya Sweeney, *Only You, Your Doctor, and Many Others May Know*, Technology Science (Sept. 2015).

[33] Yoo, Ji Su, Alexandra Thaler, Latanya Sweeney, & Jinyan Zang, *Risks to Patient Privacy: A Re-identification of Patients in Maine and Vermont Statewide Hospital Data*, Technology Science (Oct. 2018).

[34] Latanya Sweeney, Ji Su Yoo, Laura Perovich, Katherine E. Boronow, Phil Brown, & Julia Green Brody, *Re-identification Risks in HIPAA Safe Harbor Data*, Technology Science (Aug. 2017).

[35] Dr. Vanessa Teague, Dr. Chris Culnane, & Dr. Ben Rubenstein, The Simple Process of Re-identifying Patients in Public Health Records, Pursuit (Dec. 18. 2017), https://pursuit.unimelb.edu.au/articles/the-simple-process-of-re-identifying-patients-in-public-health-records.

[36] *See id.*

dataset with publicly-available information on Wikipedia and news articles of public figure women such as politicians, athletes, and celebrities.[37]

In a 2018 study on privacy vulnerabilities in anonymized California Bar Exam data, Sweeney, von Loewenfeldt, and Perry were able to re-identify individuals in spite of the presence of four anonymization protocols put forth by 'data privacy experts' in the case of *Richard Sander et. al v. State Bar of California et. al* by utilizing a host of auxiliary information -- such as online graduation programs, attorney license date data, online alumni and club membership lists.[38] Also in 2018, the public transit authority of Victoria, Australia released two billion anonymized records of travelers in Melbourne.[39] Within three months researchers Culnane, Rubinstein & Teague had confirmed re-identifications of themselves, a co-traveller and a member of the Victorian State Parliament.[40]

At this point, re-identification of "anonymized" data is taken for granted by the academic privacy community. It is no longer an open research question.

E. Reconstruction-Abetted Re-Identification Attacks Are a Realistic Threat.

Reconstruction and re-identification require neither the skills nor resources of a government agency. The Census Bureau's reconstruction and re-identification demonstration used eighteen-year-old techniques and publicly available data; the attack used no confidential data

---

[37] *Id.*

[38] Latanya Sweeney, Michael von Loewenfeldt, & Melissa Perry, *Saying It's Anonymous Doesn't Make It So: Re-identifications of "Anonymized" Law School Data*, Journal of Technology Science (Nov. 12, 2018), https://techscience.org/a/2018111301.

[39] Josh Taylor, *Myki Data Release Breached Privacy Laws and Revealed Travel Histories, Including of Victorian MP*, the Guardian (Aug. 15, 2019), https://www.theguardian.com/australia-news/2019/aug/15/myki-data-release-breached-privacy-laws-and-revealed-travel-histories-including-of-victorian-mp.

[40] *Id.*

collected by the Bureau.[41] A journalism professor was able to reproduce the reconstruction attack in "about a week."[42] Reconstruction-abetted reidentification attacks could create risks to national security. Entities who possess substantial troves of non-public personal data about the U.S. population are particularly well positioned to perform re-identification attacks on reconstructed datasets. Such non-public data might be gathered in the ordinary course of business—by Google, Facebook, Twitter, and the many data brokers that legally profit from digital surveillance—and be used to advertise, influence, and silence.

Data breaches, such as the Office of Personnel Management (OPM) hack, are another major source of non-public data that could be used by an attacker. The OPM intrusion, widely attributed to hackers working with the Chinese government, exposed detailed files and security clearance background reports on more than 21.5 million individuals.[43] The files contained both relatively mundane data such as Social Security number, date and place of birth, and in the case of security clearance background reports extremely sensitive information including data about mental health, drug use and financial problems due to gambling.[44]

Reconstruction and re-identification attacks using data from the OPM breach or any of the many other data breaches occurring in the U.S. every year, by foreign governments and others with potentially adversarial interests could create risks to national security. For example, a foreign power could undermine confidence in the Census Bureau and depress future participation in the census by using Facebook or another social media platform to reveal to 50 million Americans that

---

[41] *See* Abowd Decl. ¶ 38, Doc. 41-1.

[42] Hansen, *supra* note 12.

[43] Kim Zetter, The Massive OPM Hack Actually Hit 21 Million People, Wired (July 9, 2015 4:25 PM), https://www.wired.com/2015/07/massive-opm-hack-actually-affected-25-million/.

[44] *See id.*

their data can be reconstructed and re-identified from census responses. This could be done selectively to target particular communities. For example, it could be targeted at a particular geographic area, such as zip code, and result in selectively depressing participation.

II. Census Confidentiality Protections Must Evolve to Address Today's Threats.

A. Differential Privacy Is the Only Known Way to Protect Against Reconstruction Attacks.

Differential Privacy is the only known method for protecting large-scale statistical releases against reconstruction attacks, and hence also against reconstruction-abetted re-identification attacks. Fifteen years after its invention, there is still no effective alternative to differential privacy for defending against this threat. Given the widely understood risks described above, and unique ability of differential privacy to address them, the Census Bureau's Data Stewardship Executive Policy Committee's (DSEP) decision that the Census Bureau should use differential privacy as the core of the 2020 Disclosure Avoidance System (DAS) is wise.[45]

In addition to being the only known approach available to protect large-scale statistical releases from reconstruction attacks, differential privacy has three properties that further advance the Census Bureau's twin mandates of providing useful statistical data and protecting the confidentiality of respondents: *First*, unlike all other technologies, differential privacy is future-proof. Commercial datasets are improved and created all the time. New attacks happen all the time: Sweeney galvanized re-identification; Dinur and Nissim discovered reconstruction. Future-proofing is particularly important given the number and type of statistics the Census Bureau publishes. *Second*, differential privacy is adversary agnostic, this means it will provide protection regardless of the motivation or financial, computational, and informational assets of the adversary.

[45] *See generally* Abowd Decl. ¶ 46, Doc. 41-1.

Third, differential privacy is measurable, allowing the public to quantify cumulative privacy loss as data are analyzed and re-analyzed, shared, and linked.[46]

Moreover, systems built on differential privacy do not require secrecy of the algorithm to protect confidentiality. The ability to publicly share the implementation choices in the 2020 DAS—which the Census Bureau has publicly committed itself to do—enables stakeholders, including policy makers, data subjects and data users, to assess the level of privacy protected through those choices.[47] This creates an unprecedented increase in transparency. Stakeholders will be able to review how the agency translates its obligation to produce useful statistical reports and protect the confidentiality of participants into technical design.

This means that with differential privacy, users of the data—for example the legislatures or participants in a Voting Rights Act case—can compute confidence intervals with confidence. It means that the Census Bureau can measure privacy loss over subsequent statistical releases and censuses. It means that whether the adversary is a hostile nation state or an angry teenager the confidentiality promises the Census Bureau makes will hold. Thus, differentially private systems allow the Census Bureau and other agencies that use them to, for the first time, measure and control the total privacy loss over the huge number of statistics and statistical products it releases.

---

[46] Cynthia Dwork, Frank McSherry, Kobbi Nissim, & Adam Smith, *Calibrating Noise to Sensitivity in Private Data Analysis*, Journal of Privacy and Confidentiality 17-51 (2016), https://journalprivacyconfidentiality.org/index.php/jpc/article/download/405/388/. *See also* Cynthia Dwork, Nitin Kohli & Deirdre Mulligan, *Differential Privacy in Practice: Expose Your Epsilons!*, 9 Journal of Privacy and Confidentiality (2019), https://journalprivacyconfidentiality.org/index.php/jpc/article/view/689/685.

[47] Abowd Decl. ¶ 62 (explaining that the Census Bureau has "committed to publicly releasing the entire production code base and full suite of implementation settings and parameters"), Doc. 41-1.

B. The Census Bureau Cannot Tailor Its Confidentiality Protections to a Set of Predictable Risks as Suggested by Amicus Bambauer.

As the reconstruction attacks described above reveal, it is impossible to figure out which attacks are likely and imprudent to assume that some whole category of attack is off the table. The Census Bureau has no crystal ball: they cannot know which attacks are likely to be real threats today, let alone over the 72-year time span during which they are obligated to protect confidentiality. It is not possible for the Census Bureau to assign probabilities to attacks. The Bureau does not know what motivates the attackers, or what information (other databases) they can access. The attacker could be Facebook (yes, the whole company; nothing in the attack is illegal), for example, or employees at Facebook (or another company with enormous troves of personal data) or a company that uses a Facebook application interface (as Cambridge Analytica did) to access the vast data sources available through a social media platform like Facebooks, or a malevolent individual or organization who scrapes personal data off the web. In addition, the attacks leveraged by the Census Bureau's internal researchers, and used in the other attacks described above, are only (some of) the attack methods we are aware of today. New attack methods can, and surely will, be designed by researchers and motivated attackers.

Privacy is a non-renewable resource. If the Census Bureau assumed that a specific attack was unlikely, failed to protect against it, but then found that this attack was, in fact, going on, there would simply be no means of re-asserting additional protection. Once data with specific reconstruction or reidentification vulnerabilities has been released, they cannot be withdrawn. Protecting against worst-case attacks is the easiest way of protecting against as yet unknown realistic attacks (or attacks that will become realistic at some point in the future). It is completely misleading to characterize protection against worst-case attacks as a preoccupation with highly

unrealistic attacks (e.g., attacker who knows all but one record). The attack carried out by the Census Bureau on the 2010 release is *exactly* a case in point: it was, at the time, a new attack method that defeated the 2010 DAS protections. It was easy to carry out. No genius will be required to carry out a similar attack. In the foreseeable future, someone will likely publish or market a script or code to show others how to replicate the attack—turning yesterday's innovation into tomorrow's readily accessible weapon, usable by anyone who can download it.

C. Heuristic Alternatives Have Several Limitations.

First, they are evaluated based on outside data sources and algorithmic techniques that are available at the time. *By definition they are not "future-proof."* The Census' own internal attacks on the 2010 DAS demonstrate how fragile the guarantees can be. Consequently, such heuristics are just not reliable.

Second, heuristic approaches do not provide a measure of privacy loss. To the extent they "work"—which is typically unproven and indeed unprovable—they rely in part on secrecy. Statistical agencies are reluctant to be fully transparent about the techniques they use because adversaries can use such information to build attacks. Yet without detailed knowledge of the heuristics, it is impossible for users of the data—legislatures or participants in a Voting Rights Act case, for example—to evaluate how certain their conclusions are, by for example, computing confidence intervals.

III. Distinguishing the Census Bureau's 2020 Disclosure Avoidance System (2020 DAS) and Differential Privacy.

"Differential privacy" is a mathematical definition that some algorithms satisfy and others do not. Algorithms that satisfy the definition are called "differentially private." There are many procedures (or algorithms) that "satisfy"—meaning "adhere to"—differential privacy. For

instance, some differentially private algorithms operate by perturbing individual fields in data records, while others inject noise into the outcome of a computation or even into carefully chosen intermediate steps. A useful metaphor is to think of differential privacy as a security or safety standard that can be met in several different ways: a stopping distance standard for car brakes, for example, does not specify how the brakes should operate, but simply how quickly the vehicle must come to a stop.

Differentially private algorithms also come with a privacy budget that limits how much their output can help an attacker to make inferences about individuals in the data set. Even for a given budget, there are many different algorithms that satisfy the standard. Some will be far more accurate than others. Without familiarity with the data, it is generally difficult to determine the most accurate differentially private algorithm for a particular desired task and with a particular privacy loss budget. These kinds of questions are the subject of much research in the field.

Given this, it is crucial to distinguish discussions (critical or not) of differential privacy from discussions of the accuracy of particular algorithms or implementations. Many of the documents submitted to the court as part of this case use the terminology in confusing ways, mistaking "differential privacy" for the current proposed implementation (the proposed 2020 DAS). These include the plaintiff's original brief as well as the expert report by Dr. Barber and the amicus brief by Jane Bambauer.

IV. The 2020 DAS Does Not Use Statistical Inference.

The briefs and other materials at times use technical terms without precision to argue about the construction and application of statutory definitions. We do not offer an opinion on the correct interpretation of the statute. However, we do wish to clarify for the Court the point at which differential privacy is part of the Census Bureau’s workflow and the meaning of several terms of

art in statistics, and specifically differential privacy, that are used imprecisely, and at times inaccurately, within the record.

The DAS is applied after enumeration is completed. The DAS consists of two steps: step one uses differential privacy, and step two is a post-processing step that relies on optimization tools. Neither step involves "statistical inference" as defined in the relevant fields.

A. Differential Privacy As Used In the 2020 DAS Does Not Use Statistical Inference.

The introduction of carefully calibrated privacy-infusing random noise carried out in the first step of the 2020 DAS is most accurately viewed as "fuzzing" the details, much as faces of bystanders may be intentionally blurred out in pictures or videos. In this sense, the techniques used in the first step of the 2020 DAS are simply a more mathematically rigorous and principled alternative to the swapping than was done in the 2010 DAS, which also "fuzzed" details, typically by exchanging minority households with majority households, and which resulted in more apparent homogeneity than was actually the case. Abowd describes swapping as a form of "noise infusion."[48] No inferences are drawn, statistical or otherwise.[49]

Plaintiff's claim that "differential privacy is . . . an unlawful 'statistical method'" and that "[i]t is clear that differential privacy falls into this category" is inaccurate.[50] Barber's expert declaration, which Plaintiffs quote as support for their claims, belies their argument that differential privacy is a statistical method:

---

[48] Abowd Decl. ¶ 24, Doc. 41-1.

[49] Statistical inference is a term of art. See the definition given by Sir R. D. Cox (Oxford), inaugural winner of the International Prize in Statistics, in Appendix B.

[50] Pl.'s Motion for a Prelim. Injunction, Pt. for a Writ of Mandamus, and Mem. in Support (Mar. 11, 2021) at 38, Doc. 3.

> Privacy is introduced into the data by introducing random error through sampling from statistical distributions with parameters set to a desired level of variance . . . ***Differential privacy is thus an application of statistical processes and methods*** to adjust the original counts of the census to protect the privacy of individual records.[51]

Plaintiff is correct that the fuzzing in the 2020 DAS involves sampling from statistical distributions—usually called probability distributions. But this is not statistical inference.

B. The Post-Processing in Step Two of the DAS Does Not Use Statistical Inference.

The second step of the 2020 DAS modifies the privacy-infused statistics to satisfy certain constraints, such as consistency with state enumeration totals and certain publicly known information including the total number of housing units at the Census block level, the total number of group housing units by type in each block, and to ensure non-negative counts. The simplest analogy is to round a number to the nearest integer, for example, rounding 12.2 to 12, or 16.8 to 17 (in this rounding example, the "constraint" is that values reported are whole numbers).

Once again, there is no inference, statistical or otherwise, in this step. Michael Hawes' presentation is mistaken in using this term.[52] Hawes was referring to the use of L2 optimization, used in the 2020 DAS to perform step two as described above. Although L2 optimization can be used in statistical inference, that is not its purpose in the 2020 DAS. Plaintiff and their expert mistakenly rely on Hawes' misuse of the term.[53] The authoritative source on the approach underlying the second step in the 2020 DAS extensively describes the problem the L2 optimizer

---

[51] First Decl. of Dr. Michael Barber at 16-17, Doc. 3-5.

[52] Michael Hawes, *Differential Privacy and the 2020 Decennial Census*, U.S. Census Bureau at slide 40 (Jan. 28, 2020), https://zenodo.org/record/4122103/files/Privacy_webinar_1-28-2020.pdf.

[53] Pl.'s Reply in Support of Their Request for the Appointment of a Three-Judge Court (Mar. 25, 2021) at 1, Doc. 25; Second Expert Report of Dr. Michael Barber at 12, Doc. 25-2.

is used to address which, as described above, has nothing to do with inference – the Census Bureau already knows the confidential data![54]

Neither step of the 2020 DAS satisfies the definition of statistical inference as laid out by Sir R. D. Cox (see Appendix) and understood in the field.

CONCLUSION

The Census Bureau is tasked with providing myriad useful aggregate statistics and protecting the confidentiality of respondents. As all statistics computed from a dataset reveal small hints about the individual data records, reconstruction attacks make the Census Bureau's task more challenging. The 2010 DAS used traditional disclosure avoidance techniques that have not aged well. The Census Bureau's research, and the other well-known reconstruction attacks, document the inability of those approaches to provide any meaningful level of confidentiality today. The Census Bureau—like other statistical agencies—must adopt protections to fit the changing threats. Thanks to fifteen years of research on differential privacy, the Census Bureau has the tools to meet its statutory obligation to both provide useful statistical data and provide future-proof protection of privacy. These advances have allowed the Census Bureau to—for the very first time—measure privacy loss, fully disclose the way in which the DAS protects confidentiality, permit the computation of confidence intervals, and advance public debate about the balance between privacy and accuracy.

[54] John Abowd et al., *Census TopDown: Differentially Private Data, Incremental Schemas, and Consistency with Public Knowledge*, U.S. Census Bureau at 6 (2019), https://columbia.github.io/private-systems-class/papers/Abowd2019Census.pdf.

Respectfully submitted this 29th day of April, 2021.

/s/ Michael B. Jones
Michael B. Jones
Admitted Pro Hac Vice
Georgia Bar No. 721264
BONDURANT, MIXSON & ELMORE, LLP
1201 W. Peachtree Street, NW
Suite 3900
Atlanta, GA 30309
Telephone: (404) 881-4100
Facsimile: (404) 881-4111
Email: jones@bmelaw.com

Shannon L. Holliday (ASB-5440-Y77S)
Copeland, Franco, Screws & Gill, P.A.
P.O. Box 347
Montgomery, AL 36101-0347
Telephone: (334) 834-1180
Facsimile: (334) 834-3172
Email: holliday@copelandfranco.com

Counsel for the Data Privacy Experts

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

THE STATE OF ALABAMA, *et al.*,	)	
	)	
Plaintiffs,	)	
	)	
v.	)	Civil Action No.
	)	3:21-CV-211-RAH
UNITED STATES DEPARTMENT OF COMMERCE, *et al.*,	)	
	)	
	)	
Defendants.	)	

CERTIFICATE OF SERVICE

I hereby certify that on the 29th day of April, 2021, I electronically filed the foregoing **AMICUS BRIEF OF DATA PRIVACY EXPERTS** with the Clerk of Court using the CM-ECF system which will automatically send e-mail notification of such filing to all parties of record.

/s/ Michael B. Jones
Michael B. Jones
(Ga. Bar No. 721264)