March 13, 2018

Via electronic filing (FN-OMB-Combined-Data-RFI@omb.eop.gov)

Dr. Nancy Potok
Office of Information and Regulatory Affairs
U.S. Office of Management and Budget
725 17th St. NW
Washington, DC 20503

Re: Request for information (New techniques and methodologies based on combining data from multiple sources)

Dear Dr. Potok,

This comment is informed by research through the *Privacy Tools Project*, a broad, multidisciplinary project with collaborators across Harvard University, Georgetown University, Massachusetts Institute of Technology, Boston University, and the State University of New York at Buffalo.[1] Through this collaboration, we are exploring a wide collection of technical and legal privacy issues that arise in the context of the collection, analysis, and dissemination of research datasets containing personal information; computations over distributed data; and the publication of statistical data products by federal statistical agencies such as the US Census Bureau. Our efforts are focused on translating the theoretical promise of new measures for privacy protection such as differential privacy into practical tools and approaches. In particular, our work aims to help realize the tremendous potential from research data by making it easier to share data using privacy-protective tools.

Statistical agencies have a longstanding record of collecting, analyzing, and publishing data while protecting the privacy of respondents. Like other government and commercial actors, they are also collecting, storing, analyzing, and sharing increasingly greater quantities of personal information about individuals over progressively longer periods of time.[2] Powerful analytical capabilities, including emerging machine learning techniques, are enabling the mining of large-scale datasets to infer new insights about human characteristics and behaviors and driving demand for large-scale datasets for scientific inquiry, public policy, and innovation.

[1] Harvard University Privacy Tools Project, https://privacytools.seas.harvard.edu.

[2] See Micah Altman, Alexandra Wood, David R. O'Brien, and Urs Gasser, “Practical Approaches to Big Data Privacy Over Time,” International Data Privacy Law (forthcoming 2018), https://privacytools.seas.harvard.edu/publications/practical-approaches-big-data-privacy-over-time.

These factors are putting pressure on traditional measures for protecting privacy. Advances in the scientific study of privacy in the fields of theoretical computer science, statistics, and information science over the last two decades have demonstrated the inadequacy of widely-used privacy protection measures and other challenges related to managing information privacy in the modern world. A fundamental challenge revealed by modern privacy research is that every release of data, if it has any utility, inevitably and cumulatively, regardless of how it is protected, leaks some private information. In other words, there is no "free lunch" when using information about people; useful statistics must always be purchased with privacy loss. These advances also point to the benefits of using more recent scientifically-grounded privacy measures, as they can enable analysis of data that would have otherwise been withheld or redacted. Furthermore, such approaches can be used as tools for ensuring the validity of statistical and machine learning analyses, as they can be used to protect against overfitting.

In particular, failures of traditional privacy-preserving approaches to control disclosure risks in statistical publications have motivated computer scientists to develop a strong, formal approach to privacy. The main concept currently under study is *differential privacy*, introduced by Dwork, McSherry, Nissim, and Smith in 2006.[3] This is a formal mathematical standard for quantifying and managing privacy risk, meaning statements about risk are proved mathematically (rather than, say, empirically). The definition requires the output distribution of a privacy preserving analysis to remain "stable" under any possible change to a single individual's information. Currently, differential privacy is the only framework that provides meaningful privacy guarantees in scenarios in which adversaries have access to arbitrary external information. Analyses satisfying differential privacy provide provable privacy protection against any feasible adversarial attack, whereas de-identification concepts only counter a limited set of specific attacks.[4]

Differential privacy has a compelling intuitive interpretation as it essentially masks the contribution of any single individual, making it impossible to infer any information specific to him or her, including whether the individual's information was used at all. It can also be interpreted as essentially ensuring that using an individual's data will not reveal any personally identifiable information that is specific to him or her, i.e., information that cannot be inferred unless the individual's information is used in the analysis.

[3] See Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," Journal of Privacy and Confidentiality 7(3): 2 (2016), http://repository.cmu.edu/jpc/vol7/iss3/2/.

[4] For a more detailed but intuitive discussion of the protection provided by differential privacy, Kobbi Nissim, Thomas Steinke, Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, David O'Brien, and Salil Vadhan, "Differential Privacy: A Primer for a Non-technical Audience," Working Paper (2018), https://privacytools.seas.harvard.edu/publications/differential-privacy-primer-non-technical-audience-preliminary-version.

There is a continually growing list of tasks that have been shown, in principle, to be computable with differential privacy including descriptive and inferential statistics, machine learning algorithms, and production of synthetic data. Existing real-world applications of differentially private analyses include implementations by companies such as Google, Apple, and Uber,[5] and federal agencies such as the US Census Bureau.[6]

Differentially private tools can help enable researchers, policymakers, and businesses to analyze and share sensitive data while providing strong guarantees of privacy to the individuals in the data. Differential privacy is supported by a rich and rapidly advancing theory that enables one to reason with mathematical rigor about privacy risk. Adopting this formal approach to privacy yields a number of practical benefits for users:[7]

- Systems that adhere to strong formal definitions like differential privacy provide protection that is robust to a wide range of potential privacy attacks, as defined above, including attacks that are unknown at the time of deployment. An analyst designing a differentially private data release need not anticipate particular types of privacy attacks, such as the likelihood that one could link particular fields with other data sources that may be available.

- Differential privacy provides provable privacy guarantees with respect to the cumulative risk from successive data releases and is the only existing approach to privacy that provides such a guarantee

- Differentially private tools also have the benefit of transparency, as it is not necessary to maintain secrecy around a differentially private computation or its parameters. This feature distinguishes differentially private tools from traditional de-identification techniques which often require concealment of the extent to which the data have been transformed, thereby leaving data users with uncertainty regarding the accuracy of

[5] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. "Rappor: Randomized aggregatable privacy-preserving ordinal response." In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054-1067. ACM, 2014; Greenberg, Andy. "Apple's 'differential privacy' is about collecting your data–but not your data." *Wired (June 13, 2016)*(2016); Johnson, Noah, Joseph P. Near, and Dawn Song. "Practical differential privacy for sql queries using elastic sensitivity." *arXiv preprint arXiv:1706.09479* (2017).

[6] National Academies of Sciences, Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps, Consensus Report (2017), https://www.nap.edu/catalog/24893/federal-statistics-multiple-data-sources-and-privacy-protection-next-steps; John M. Abowd, "Why the Census Bureau Adopted Differential Privacy for the 2020 Census of Population," Presentation at Harvard University (December 11, 2017), https://privacytools.seas.harvard.edu/why-census-bureau-adopted-differential-privacy-2020-census-population.

[7] See Kobbi Nissim, Thomas Steinke, Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, David O'Brien, and Salil Vadhan, "Differential Privacy: A Primer for a Non-technical Audience," Working Paper (2018), https://privacytools.seas.harvard.edu/publications/differential-privacy-primer-non-technical-audience-preliminary-version.

analyses on the data. This can enable public scrutiny of the privacy-preserving techniques used.

- Differentially private tools can be used to provide broad, public access to data or data summaries in a privacy-preserving way. They can enable wide access to data that cannot otherwise be shared due to privacy concerns, and do so with a guarantee of privacy protection that substantially increases the ability of the institution to protect the individuals in the data.

Furthermore, data need not be centralized to perform a differentially private analysis. The field of cryptography offers formal approaches to performing computations on joint data while not leaking more information than the functionality intended. For example, secure multiparty computation can be used to perform differentially private computations over distributed data.[8]

Regulations and policies for privacy protection should evolve in light of scientific advances in privacy. In particular, expectations that traditional disclosure control techniques such as de-identification provide sufficient privacy protection are no longer supported by the legal or scientific literature.[9] A key insight from the scientific study of privacy is that data cannot be analyzed or released without some leakage of information about individuals. Differential privacy quantifies this leakage and, furthermore, is equipped with tools for bounding the accumulation of multiple releases. It is a matter of policy to set a limit for privacy leakage (referred to as the "privacy budget") and decide how to act once the budget is exhausted. Policymakers should accordingly consider the importance of setting and monitoring a privacy budget and develop policies specifying how the privacy budget should be used, such as how to choose between analyses to be performed if the privacy budget cannot allow all desired analyses.[10]

Differential privacy is a new way of protecting privacy that is more quantifiable and comprehensive than the concepts of privacy that underlie many existing laws, policies, and practices around privacy and data protection. The differential privacy guarantee can be interpreted in reference to these other concepts, and can even accommodate variations in how they are defined across different laws. In many cases, data holders may use differential privacy

[8] For an early example of this approach, see Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor, "Our Data, Ourselves: Privacy via Distributed Noise Generation," *Advances in Cryptology - EUROCRYPT 2006* (2006): 486–503, https://www.iacr.org/archive/eurocrypt2006/40040493/40040493.pdf.

[9] See Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Review* 57 (2010): 1701-1777.

[10] For an intuitive understanding of the privacy loss parameter and privacy budget, see Kobbi Nissim, Thomas Steinke, Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, David O'Brien, and Salil Vadhan, "Differential Privacy: A Primer for a Non-technical Audience," Working Paper (2018), https://privacytools.seas.harvard.edu/publications/differential-privacy-primer-non-technical-audience-preliminary-version.

to demonstrate that they have complied with legal and policy requirements for privacy protection.

However, the diversity of regulatory standards for privacy protection introduces ambiguity and uncertainty with respect to the level of protection needed when data governed by different regulatory and policy regimes are combined. Consequently, it is important to work towards unifying the regulatory requirements and harmonize them with the scientific knowledge being accumulated about privacy. Increasing the understanding of how formal privacy models relate to basic concepts from privacy law (such as personally identifiable information, linkage, identification, risk, and inference) can help reduce this uncertainty. An example of work in this direction is research by Nissim et al. on bridging the gaps between differential privacy and the de-identification requirements of the Family Educational Rights and Privacy Act (FERPA).[11] This approach involved (i) identifying a description of a potential privacy attacker and the attacker's goals embedded within FERPA's definition of personally identifiable information; (ii) developing a mathematical model of FERPA's privacy requirements while making conservative assumptions in the model to account for possible ambiguities in the regulatory standard; and (iii) analyzing differential privacy with respect to the mathematical model extracted from FERPA's privacy requirements, and proving formally that differential privacy is sufficient to satisfy its requirements.

In other work, we have argued for the need for comprehensive and consistent regulatory protection against information privacy harms in research.[12] Protection for people whose information is used in research should be based on the risks and benefits to the subject and to society, and not on other elements of the research context that are irrelevant from an ethical perspective, such as the institution conducting the research, its commercial status, or its sources of funding.

There is a growing need to develop and implement approaches to privacy that are up-to-date scientifically and to combine these rigorous approaches with additional legal and procedural

[11] See Kobbi Nissim, Aaron Bembenek, Alexandra Wood, Mark Bun, Marco Gaboardi, Urs Gasser, David R. O'Brien, Thomas Steinke, and Salil Vadhan, "Bridging the Gap between Computer Science and Legal Approaches to Privacy," 31 *Harvard Journal of Law & Technology* __ (forthcoming 2018), https://privacytools.seas.harvard.edu/publications/bridging-gap-between-computer-science-and-legal-approaches-privacy; Alexandra Wood, "Bridging Privacy Definitions: Differential Privacy and Concepts from Census Law & Policy," Presentation at DIMACS/Northeast Big Data Hub Workshop on Overcoming Barriers to Data Sharing, Rutgers University in New Brunswick, NJ (Oct. 23-24, 2017), http://dimacs.rutgers.edu/Workshops/Barriers/Slides/Wood.pdf.

[12] See Effy Vayena, Urs Gasser, Alexandra Wood, David R. O'Brien, and Micah Altman, "Elements of a New Ethical Framework for Big Data Research," *Washington and Lee Law Review* 72(3) (2016): 420-441, http://lawreview.journals.wlu.io/elements-of-a-new-ethical-framework-for-big-data-research.

tools. As a general framework, we recommend that policies and guidance on privacy and confidentiality be based on the following principles of a modern approach to privacy:[13]

- Calibrating privacy and security controls to the intended uses and privacy risks associated with the data;
- When conceptualizing informational risks, considering not just re-identification risks but also inference risks, or the potential for others to learn about individuals from the inclusion of their information in the data;
- Addressing informational risks using a combination of privacy and security controls rather than relying on a single control such as consent or deidentification;
- Calibrating controls to the specific structure and risks of the data as datasets increase in scale, heterogeneity, integration and longevity, both the benefits of data use and privacy risks may grow non-linearly (e.g., the protection devised for individual uses of datasets does not necessarily protect the combined uses);
- Anticipating, regulating, monitoring, and reviewing interactions with data across all stages of the lifecycle (including the postaccess stages), as risks and methods will evolve over time; and
- In efforts to harmonize approaches across regulations and institutional policies, emphasizing the need to provide similar levels of protection to research activities that pose similar risks.

In addition, the collection and use of fine-grained personal data over time is associated with significant risks to individuals, groups, and society at large. The risks posed by big data are a function of temporal factors comprising age, period, and frequency and non-temporal factors such as population diversity, sample size, dimensionality, and intended analytic use. Increasing complexity in any of these factors, individually or in combination, creates heightened risks that are not readily addressable through traditional de-identification and process controls. However these risks can be mitigated by new privacy technologies approaches when these are calibrated to risk factors present in a specific case.[14]

[13] See Micah Altman, Alexandra Wood, David O'Brien, Salil Vadhan, and Urs Gasser, “Towards a Modern Approach to Privacy-Aware Government Data Releases,” Berkeley Technology Law Journal 30(3): 1967-2072 (2015), http://btlj.org/data/articles2015/vol30/30_3/1967-2072%20Altman.pdf; Micah Altman, Alexandra Wood, David R. O'Brien, and Urs Gasser, “Practical Approaches to Big Data Privacy Over Time,” International Data Privacy Law (forthcoming 2018), https://privacytools.seas.harvard.edu/publications/practical-approaches-big-data-privacy-over-time.

[14] See Micah Altman, Alexandra Wood, David R. O'Brien, and Urs Gasser, “Practical Approaches to Big Data Privacy Over Time,” *International Data Privacy Law* (forthcoming 2018), https://privacytools.seas.harvard.edu/publications/practical-approaches-big-data-privacy-over-time.

Sincerely,

Micah Altman
MIT Libraries, Massachusetts Institute of Technology

Aloni Cohen
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

Aaron Fluitt
Georgetown University Law Center

James Honaker
Center for Research on Computation and Society, Harvard University

Kobbi Nissim
Department of Computer Science, Georgetown University

Michael Washington
J.D. Candidate, 2018, Washington University in St. Louis

Alexandra Wood
Berkman Klein Center for Internet & Society, Harvard University

Citations

John M. Abowd, "Why the Census Bureau Adopted Differential Privacy for the 2020 Census of Population," Presentation at Harvard University (December 11, 2017), https://privacytools.seas.harvard.edu/why-census-bureau-adopted-differential-privacy-2020-census-population.

Micah Altman, Alexandra Wood, David R. O'Brien, and Urs Gasser, "Practical Approaches to Big Data Privacy Over Time," *International Data Privacy Law* (forthcoming 2018), https://privacytools.seas.harvard.edu/publications/practical-approaches-big-data-privacy-over-time.

Micah Altman, Alexandra Wood, David O'Brien, Salil Vadhan, and Urs Gasser, "Towards a Modern Approach to Privacy-Aware Government Data Releases," *Berkeley Technology Law*

Journal 30(3): 1967-2072 (2015), http://btlj.org/data/articles2015/vol30/30_3/1967-2072%20Altman.pdf.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor, “Our Data, Ourselves: Privacy via Distributed Noise Generation,” *Advances in Cryptology - EUROCRYPT 2006* (2006): 486–503, https://www.iacr.org/archive/eurocrypt2006/40040493/40040493.pdf.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith, “Calibrating Noise to Sensitivity in Private Data Analysis,” *Journal of Privacy and Confidentiality* 7(3): 2 (2016), http://repository.cmu.edu/jpc/vol7/iss3/2/.

National Academies of Sciences, Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps, Consensus Report (2017), https://www.nap.edu/catalog/24893/federal-statistics-multiple-data-sources-and-privacy-protection-next-steps.

Kobbi Nissim, Thomas Steinke, Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, David O'Brien, and Salil Vadhan, “Differential Privacy: A Primer for a Non-technical Audience,” Working Paper (2018), https://privacytools.seas.harvard.edu/publications/differential-privacy-primer-non-technical-audience-preliminary-version.

Kobbi Nissim, Aaron Bembenek, Alexandra Wood, Mark Bun, Marco Gaboardi, Urs Gasser, David R. O'Brien, Thomas Steinke, and Salil Vadhan, “Bridging the Gap between Computer Science and Legal Approaches to Privacy,” *Harvard Journal of Law & Technology* (forthcoming 2018), https://privacytools.seas.harvard.edu/publications/bridging-gap-between-computer-science-and-legal-approaches-privacy.

Alexandra Wood, "Bridging Privacy Definitions: Differential Privacy and Concepts from Census Law & Policy," Presentation at DIMACS/Northeast Big Data Hub Workshop on Overcoming Barriers to Data Sharing, Rutgers University in New Brunswick, NJ (Oct. 23-24, 2017), http://dimacs.rutgers.edu/Workshops/Barriers/Slides/Wood.pdf.