BOSTON TECH HUB FACULTY WORKING GROUP

SPRING SESSION 1 • APRIL 21, 2020

Differential Privacy

PREPARED BY:

Mike Miesen
Research Assistant, Belfer Center for Science and International Affairs

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs
TECHNOLOGY & PUBLIC PURPOSE PROJECT

Harvard John A. Paulson
School of Engineering
and Applied Sciences

The Boston Tech Hub Faculty Working Group, hosted by former Secretary of Defense and Harvard Kennedy School Belfer Center Director Ash Carter and Harvard SEAS Dean Frank Doyle, will convene its first session of the spring semester on the topic of differential privacy. The session will examine data privacy issues, current applications of differential privacy tools, capabilities, limitations, and more.

PROBLEM TO BE ADDRESSED

Governments, businesses, and academics rely on aggregated data to understand problems, test hypotheses, and improve operations. However, even aggregated, de-identified personal data can be individually re-identified through myriad known and currently-unknown methods, which erodes individual security and privacy. **Striking an appropriate balance between societal value stemming from the use of data with individual privacy and security is of paramount concern**.

Context

- **Differential privacy** is a "formal mathematical framework for quantifying and managing privacy risks."[1] and "a general framework for reasoning about the increased risk that is incurred when an individual's information is included in a data analysis."[2]

- Tactically, differential privacy tools add a calibrated amount of statistical noise to datasets, depending on the value of a privacy loss parameter (epsilon, or the "privacy budget"). Small values of epsilon denote high amounts of added statistical noise, limiting the usefulness of the analysis but increasing privacy protection. As a result, differential privacy tools create analyses that are approximations of "real-world" analyses, trading off perfectly-accurate analyses for increased individual privacy protection.

- Importantly, researchers believe differential privacy tools can "future-proof" statistical analyses; that is, they can give researchers and individuals confidence that future technologies—advanced computing power, new algorithms, and the like—will not re-identify individuals' data in what is known as a *privacy attack*.

[1] Alexandra Wood et al. "Differential Privacy: A Primer for a Non-Technical Audience." February 2019. Berkman Klein Center for Internet and Society at Harvard University. Page 209

[2] Ibid, 240.

Applications

- **2020 United States Census.** The 2020 US Census will incorporate differential privacy tools to protect public information. As the Census Bureau notes, "There are many variants of differential privacy. The one selected for the 2020 Census introduces controlled noise into the data in a manner that preserves the accuracy at higher levels of geography…Our differential privacy methods will be designed to preserve the utility of our legally mandated data products while also ensuring that every respondents' personal information is fully protected."[3]

- **Private Sector.** Companies use differential privacy tools to gather aggregated user activity data to improve their services for users. Apple, for example, uses differential privacy techniques to collect and analyze data on emoji use "to help design better ways to find and use our favorite emoji."[4]

Public Purpose Concerns and Considerations

- **Data Privacy.** With increasing processing power and available data, reidentifyingdata that has been deidentified will become easier over time.

  For example, in 2006, Netflix released deidentified user data as part of its 'Netflix Prize' competition, which researchers were able to reidentify using other public data.[5] US Census Bureau researchers were able to conduct successful 'privacy attacks' on previous census data, too: "when Census Bureau researchers accounted for modern algorithms and computing power, they discovered the inadequacy of these measures. Like with Netflix, security through obscurity collapsed when other public data sources were combined with the last census."[6]

  Differential privacy tools quantify the potential risk of data privacy loss for individuals, creating an upper-bound on potential privacy loss.

- **Data Accuracy & Sample Size.** While all statistical analyses have inherent inaccuracies—stemming from, among other things, sampling error—differentially private analyses add another type of error: "Analyses performed with differential privacy differ from standard statistical analyses—such as the

[3] Jarmin, Ron. "Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census." United States Census Bureau. February 15th, 2019. https://www.census.gov/newsroom/blogs/random-samplings/2019/02/census_bureau_adopts.html

[4] "Differential Privacy." Apple. Page 3. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

[5] Francis, Matthew. "Using Differential Privacy to Protect the United States Census." *Siam News*. Society for Industrial and Applied Mathematics. October 1st, 2019. https://sinews.siam.org/Details-Page/using-differential-privacy-to-protect-the-united-states-census

[6] Ibid.

calculation of averages, medians, and linear regression equations—in that random noise is added in the computation."[7]

Governance and Regulation

- Domestically and internationally, several laws mandate that governments, organizations, and researchers to protect data privacy (e.g., Federal Policy for the Protection of Human Subjects, General Data Protection Regulations), with additional protection for certain sensitive data (e.g., health, education).[8] As Wood et al put it: Taken together, the safeguards required by these legal and ethical frameworks are designed to protect the privacy of individuals and ensure they fully understand both the scope of personal information to be collected and the associated privacy risks. They also help data holders avoid administrative, civil, and criminal penalties, as well as maintain the public's trust and confidence in commercial, government, and research activities involving personal data."[9]

- Differential privacy is viewed as a tool that can help satisfy legal requirements: "Interest in the concept is growing among potential users of the tools, as well as within legal and policy communities, as it holds promise as a potential approach to satisfying legal requirements for privacy protection when handling personal information. In particular, differential privacy may be seen as a technical solution for analyzing and sharing data while protecting the privacy of individuals in accordance with existing legal or policy requirements for de-identification or disclosure limitation."[10]

[7] Alexandra Wood et al. "Differential Privacy: A Primer for a Non-Technical Audience." February 2019. Berkman Klein Center for Internet and Society at Harvard University. Page 220.

[8] Ibid.

[9] Ibid, 216.

[10] Ibid, 210.

Discussion Questions

- How does differential privacy compare as a solution to other means of protecting data privacy (e.g., homomorphic encryption, k-anonymization)
- What are the tradeoffs of using differential privacy tools to protect data privacy?
- How can differential privacy tools scale beyond traditional/anticipated use cases, if at all?

Readings

Alexandra Wood, Salil Vadhan et al. "Differential Privacy: A Primer for a Non-Technical Audience." 2019.

Mark Bun. "A Teaser for Differential Privacy." Princeton. December 2017.

Michael Hawes. "Differential Privacy, and the 2020 Decennial Census." United States Census Bureau. March 2020.