July 8, 2021

Differential Privacy in the Real World

An introduction and overview of differential privacy

Claire McKay Bowen, Ph.D
Lead Data Scientist, Privacy and Data Security
cbowen@urban.org www.clairemckaybowen.com @ClaireMKBowen

1

Overview

1. **Motivation:** What is data privacy? Why should we care?
2. **Background:** What has been done to address these issues?
3. **Methodology:** What is differentially private data synthesis?
4. **Challenges:** What are the ongoing challenges in the field?
5. **Future Work**

2

BRIEFING ROOM

Executive Order On Advancing Racial Equity and Support for Underserved Communities Through the Federal Government

JANUARY 20, 2021 • PRESIDENTIAL ACTIONS

"…advanc[e] equity for all, including people of color and others who have been historically underserved, marginalized, and adversely affected by persistent poverty and inequality."

3

Lack of Public Data Hampers COVID-19 Fight

STATELINE ARTICLE August 3, 2020 By: Christine Vestal Topics: Health Read time: 7 min

A ventilator helps a COVID-19 patient breathe at a Houston hospital. Hospital data related to the coronavirus pandemic will now be collected by a private technology firm, rather than the Centers for Disease Control and Prevention. Epidemiologists say better COVID-19 data is needed to improve the nation's response.

David J. Phillip/The Associated Press

Google Promises Privacy With Virus App but Can Still Collect Location Data

Some government agencies that use the software said they were surprised that Google may pick up the locations of certain app users. Others said they had unsuccessfully pushed Google to make a change.

Switzerland has asked Google to decouple the location setting requirement on Android phones from Bluetooth, which the country's virus alert app uses to detect nearby smartphones. Fabrice Coffrini/Agence France-Presse — Getty Images

By Natasha Singer

July 20, 2020

URBAN INSTITUTE

3

4

HEALTH

'It's not a pretty picture': Why the lack of racial data around COVID vaccines is 'massive barrier' to better distribution

Nada Hassanein USA TODAY

Published 5:30 a.m. ET Feb. 1, 2021 | Updated 2:10 p.m. ET Feb. 1, 2021

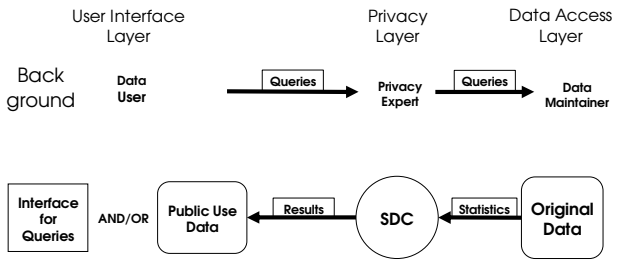
Abigail Echo-Hawk, chief research officer with Seattle Indian Health Board and a member of the Pawnee Tribe, gets a shot of the Moderna COVID-19 vaccine on Dec. 21. A colleague used a black pen to inscribe "For the (Heart) love of Native People" over the injection spot. *Karen Ducey, Getty Images*

URBAN INSTITUTE

4

5

How do we access confidential data?

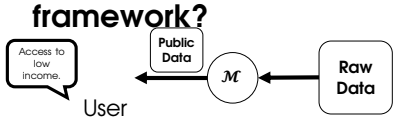


URBAN INSTITUTE

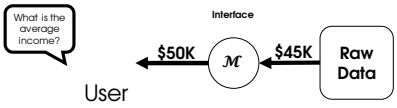
6

6

What are the issues with this framework?



How do we **measure utility** (usefulness) and **disclosure risk** of the data?



How **much noise should be added** and how do you **limit the number of queries**?

URBAN INSTITUTE 7

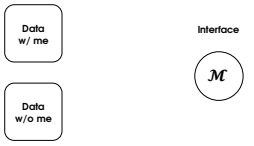
7

What is differential privacy?

A sanitization algorithm, $\mathcal{M}$, is ϵ-differentially private if for all subsets $S \subseteq Range(\mathcal{M})$ and for all X, X' such that $d(X, X') = 1$,

$$\frac{\Pr(\mathcal{M}(X) \in S)}{\Pr(\mathcal{M}(X') \in S)} \leq \exp(\epsilon)$$

where $\epsilon > 0$ is the privacy-loss budget and $d(X, X') = 1$ represents the possible ways that X' differs from X.



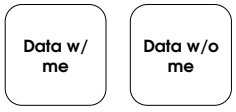
URBAN INSTITUTE Dwork, McSherry, Nissim, Smith (2006) 8

8

What is the Global Sensitivity?

Global (L_1) Sensitivity: For all X, X' such that $d(X, X') = 1$, the global sensitivity of a function u is

$$\Delta_1 u = \sup_{d(X,X')=1} \left|\left|u(X) - u(X')\right|\right|_1$$



URBAN INSTITUTE Dwork, McSherry, Nissim, Smith (2006) 9

9

What is the Laplace mechanism?

Proposition 1: The Laplace mechanism satisfies ϵ-differentially private by adding noise to query, u, such that

$$u^*(X) = u(X) + Lap\left(0, \frac{\Delta_1 u}{\epsilon}\right)$$

where $\Delta_1 u$ is the l_1-global sensitivity.

URBAN INSTITUTE Dwork, McSherry, Nissim, Smith (2006) 10

10

What is the Laplace mechanism?

Proposition 1: The Laplace mechanism satisfies ϵ-differentially private by adding noise to query, u, such that

$$u^*(X) = u(X) + Lap\left(0, \frac{\Delta_1 u}{\epsilon}\right)$$

where $\Delta_1 u$ is the l_1-global sensitivity.

URBAN INSTITUTE Dwork, McSherry, Nissim, Smith (2006) 10

11

What are the important differential privacy theorems?

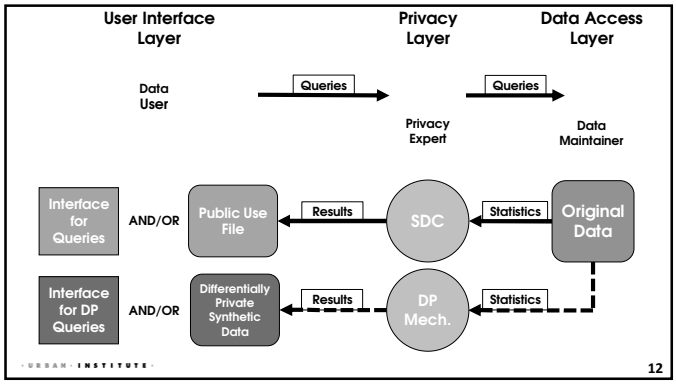
Composition Theorems: Suppose mechanism, $\mathcal{M}$, provides ϵ-differential privacy for $j = 1, \ldots, k$.

1. **Sequential Composition:** The sequence of $\mathcal{M}_j(X)$ applied on the same X provides $\sum_j \epsilon_j$-differential privacy.
2. **Sequential Composition:** Let D_j be disjoint subsets of the input domain D. The sequence of $\mathcal{M}_j(X \cap D_j)$ provides $\max(\epsilon_j)$-differential privacy.

Post-Processing Theorem: If $\mathcal{M}$ be a mechanism that satisfies ϵ-differential privacy and g be any function, then $g(\mathcal{M}(X))$ also satisfies ϵ-differential privacy

URBAN INSTITUTE Dwork, McSherry, Nissim, Smith (2006) 11

12



13

What is non-parametric data synthesis?

Marginal Tables – a basic synthetic data approach by using random sampling

URBAN INSTITUTE Dwork et. al (2006), Wasserman & Zhou (2010) 13

14

What is non-parametric data synthesis?

Marginal Tables – a basic synthetic data approach by **using random sampling**

URBAN INSTITUTE Dwork et. al (2006), Wasserman & Zhou (2010) 13

15

What is non-parametric data synthesis + DP?

Marginal Tables – a basic synthetic data approach by **using random sampling** + additional noise from a **differentially private mechanism**

 Dwork et. al (2006), Wasserman & Zhou (2010)

16

What is parametric data synthesis?

Probability Distribution – identify an appropriate probability distribution and make draws from that distribution

 Wasserman & Zhou (2010), Bowen and Liu (2020)

17

What is parametric data synthesis?

Probability Distribution – identify an appropriate probability distribution and **make draws from that distribution**

 Wasserman & Zhou (2010), Bowen and Liu (2020)

18

What is parametric data synthesis + DP?

Probability Distribution – identify an appropriate probability distribution and **make draws from that distribution** + additional noise from a **differentially private mechanism** to the parameters (sufficient statistics)

URBAN INSTITUTE Wasserman & Zhou (2010), Bowen and Liu (2020) 14

19

What is are the issues with differentially private data synthesis?

Non-Parametric

- does **not preserve** the variability in the data
- bias issues with **sparsity**

Parametric

- is **model dependent**
- global sensitivity can be **difficult to calculate** without making assumptions on bounds
- is **not scalable** with increased parameters

URBAN INSTITUTE Bowen and Liu (2020) 15

20

NIST
Communications Technology Laboratory
PUBLIC SAFETY COMMUNICATIONS RESEARCH DIVISION

2018 Differential Privacy Synthetic Data Challenge

Challenge Details

The Differential Privacy Synthetic Data Challenge tasked participants with creating new methods, or improving existing methods of data de-identification, while preserving the dataset's utility for analysis. Competitors participated in three marathon matches on the Topcoder platform with the goal of designing, implementing, and proving that their synthetic data generation algorithm satisfied differential privacy. All solutions were required to satisfy the differential privacy guarantee, a provable guarantee of individual privacy protection.

2020 Differential Privacy Temporal Map Challenge

Differential Privacy Temporal Map Challenge

Visit the Challenge Website.

Visit Challenge.gov for the Official Rules

Open for submissions October 1, 2020 - May 17, 2021

URBAN INSTITUTE 16

21

How do move past non-tabular data?

Google COVID-19 Community Mobility Reports

See how your community is moving around differently due to COVID-19

As global communities respond to COVID-19, we've heard from public health officials that the same type of aggregated, anonymized insights we use in products such as Google Maps could be helpful as they make critical decisions to combat COVID-19.

These Community Mobility Reports aim to provide insights into what has changed in response to policies aimed at combating COVID-19. The reports chart movement trends over time by geography, across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential.

 Aktay et al. (2020)

22

How do we set the privacy loss budget?

FOR IMMEDIATE RELEASE: WEDNESDAY, JUNE 09, 2021

Census Bureau Sets Key Parameters to Protect Privacy in 2020 Census Results

JUNE 09, 2021
RELEASE NUMBER CB21-CN.42

SUBSCRIBE RSS SMS EMAIL

JUNE 9, 2021 –The U.S. Census Bureau's Data Stewardship Executive Policy Committee (DSEP) announced it has selected the settings and parameters for the Disclosure Avoidance System (DAS) for the 2020 Census redistricting data (PL-94-171). The DAS uses a mathematical algorithm to ensure that the privacy of individuals is sufficiently protected while maintaining high levels of accuracy in the statistics we produce.

The Census Bureau released the first "beta" version of the DAS in October 2019, and released further demonstration data products in May, September, and November 2020, and in April 2021. During this process, independent experts and stakeholders, along with data users, have provided extensive feedback to help shape each subsequent test product and to inform the decisions.

23

How do we set the privacy loss budget?

Use Case	Privacy Model	DP Algorithm Parameters (ϵ, δ)	Daily DP Parameters $(\epsilon_{\text{day}}, \delta_{\text{day}})$	Monthly DP Parameters $(\epsilon_{\text{month}}, \delta_{\text{month}})$
Google - RAPPOR [3] Chrome Homepages	Local[b]	$(0.534, 0)$	$(25.63, 0)$ 30 min reporting	$(769, 0)$
Apple - Safari Domains [4]	Local	$(4, 0)$	$(8, 0)^a$	$(240, 0)$
Apple - Emojis [4]	Local	$(4, 0)$	$(4, 0)^a$	$(120, 0)$
Microsoft - Telemetry Collection per App [5]	Local[b]	$(0.686, 0)$	$(2.74, 0)$ 6 hour reporting	$(82.2, 0)$
Google - Mobility Reports [24]	Global	$(0.11, 0)$ or $(0.22, 0)$	$(2.64, 0)^c$	$(79.2, 0)$
Microsoft - Assistive AI[d]	Global	$(4, 10^{-7})$	Not available	Not available
LinkedIn - Audience Engagement API[e]	Global	$(0.15, 10^{-10})$	—	$(34.9, 7 \times 10^{-9})$

Table 2: Privacy parameters for existing deployments of privacy systems that use differentially private algorithms. Note that some parameters can be improved with a slightly larger δ_{month}.

 Rogers et al. (2020)

24

HARVARD UNIVERSITY HARVARD.EDU

OpenDP

building an open-source suite of tools for deploying differential privacy

Home About People Events Blog

We are engaging a community of collaborators in academia, industry, and government to build trustworthy, open-source software tools for privacy-protective statistical analysis of sensitive personal data. These tools, which we call OpenDP, will offer the rigorous protections of differential privacy for the individuals who may be represented in confidential data and statistically valid methods of analysis for researchers who study the data.

We began this project in partnership with Microsoft developing a differentially private data curator application. Building on this collaboration, we are now building a broader community around OpenDP with stakeholders and contributors from across academia, industry, and government. Together, we plan to design, implement, and govern an "OpenDP Commons" that includes a library of differentially private algorithms and other general-purpose tools for use in end-to-end differential privacy systems.

OpenDP is being incubated by Harvard University's Privacy Tools and Privacy Insights projects (at SEAS and IQSS), with generous support from the Sloan Foundation.

Visit us on Github: https://github.com/opendifferentialprivacy/

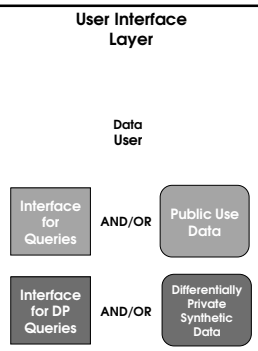
23

25

What are other ways to access data?

The goal of developing a **validation server** is to significantly **expand access to restricted and confidential government data** for researchers and analysts, while ensuring individual privacy and complying with all applicable regulations.

- Tier I – Public Use Data
- Tier II – Validation Server
- Tier III – Confidential Data



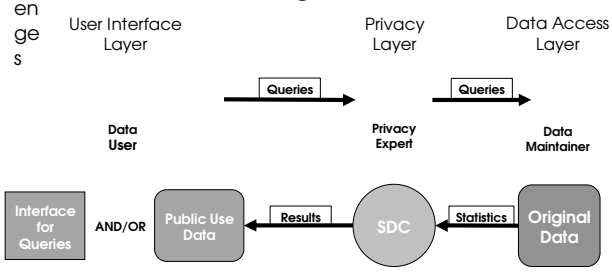
URBAN INSTITUTE

25

26

• Ch
d
en
ge
s

What are the changes to data access?



URBAN INSTITUTE

27

27

Summary

- Data Privacy
- Data Synthesis
- Differential Privacy
- Non-Parametric & Parametric Differentially Private Data Synthesis
- Challenges
  - Scientific Challenges
  - Transition Challenges
  - Social Challenges

Contact Me

cbowen@urban.org

www.clairemckaybowen.com

/in/bowenclaire

@ClaireMKBowen

URBAN INSTITUTE

28

References

1. Aktay, A., Bavadekar, S., Cossoul, G., Davis, J., Desfontaines, D., Fabrikant, A., ... & Wilson, R. J. (2020). Google COVID-19 Community Mobility Reports: anonymization process description (version 1.1). *arXiv preprint arXiv:2004.04145*.
2. Bowen, CMK. & Liu, F. (2021). Comparative study of differentially private data synthesis methods. *Statistical Science, 35*(2), 280-307.
3. Bowen, C. M., & Snoke, J. (2021). Comparative Study of Differentially Private Synthetic Data Algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge. *Journal of Privacy and Confidentiality*, *11*(1).
4. Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography (pp. 265-284). *Springer Berlin Heidelberg*.
5. Rogers, R., Subramaniam, S., Peng, S., Durfee, D., Lee, S., Kancha, S. K., ... & Ahammad, P. (2020). LinkedIn's Audience Engagements API: A privacy preserving data analytics system at scale. *arXiv preprint arXiv:2002.05839*.
6. Wasserman, L., & Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489), 375-389.

URBAN INSTITUTE

29