Data and Society
Differential Privacy – Lecture 5

2/8/21

Today (2/8/21)

- Lecture / Discussion – Differential Privacy
- Student Presentations
- The next assignment (given February 11) will be a Briefing. You will do this in teams of 2.
  - **If you would like to pick your own partner, send an email (ccing both partners) to FB by 2/9** letting me know who is on your team
  - If you don't pick your own partner (which is fine …), FB will assign you one.
  - Instructions for the Briefing will be given next time.

Reading for February 11

- **"The Fantasy of Opting Out"**, MIT Press Reader, https://thereader.mitpress.mit.edu/the-fantasy-of-opting-out/

The Fantasy of Opting Out

Those who know about us have power over us. Obfuscation may be our best digital weapon.

There are still ways to carve out spaces of resistance, counterargument, and autonomy. Source image: Lianhao Qu, via Unsplash

Date	Topic	Speaker	Date	Topic	Speaker
1-25	Introduction	Fran	1-28	The Data-driven World	Fran
2-1	Data and COVID-19	Fran	2-4	Data and Privacy -- Intro	Fran
2-8	Data and Privacy – Differential Privacy	Fran	2-11	Data and Privacy – Anonymity	Fran
2-15	NO CLASS / PRESIDENT'S DAY		2-18	Data and Privacy – Law	Ben Wizner
2-22	Digital rights in the EU and China	Fran	2-25	Data and Discrimination 1	Fran
3-1	Data and Discrimination 2	Fran	3-4	Data and Elections 1	Fran
3-8	Data and Elections 2	Fran	3-11	NO CLASS / WRITING DAY	
3-15	Data and Astronomy	Alyssa Goodman	3-18	Data Science	Fran
3-22	Digital Humanities	Brett Bobley	3-25	Data Stewardship and Preservation	Fran
3-29	Data and the IoT	Fran	4-1	Data and Smart Farms	Rich Wolski
4-5	Data and Self-Driving Cars	Fran	4-8	Data and Ethics 1	Fran
4-12	Data and Ethics 2	Fran	4-15	Cybersecurity	Fran
4-19	Data and Dating	Fran	4-22	Data and Social Media	Fran
4-26	Tech in the News	Fran	4-29	Wrap-up / Discussion	Fran
5-3	NO CLASS				

Lecture – Differential Privacy

- What is differential privacy?
- Differential Privacy and the Census

What is Differential Privacy and Why is it useful?

- What is Differential Privacy? NIST
- https://www.youtube.com/watch?v=-JRURYTfBXQ

Differential Privacy [Dwork et al.]

- Differential Privacy is an approach which promotes **privacy in statistical databases**, i.e. information about the **group** but not specific individuals
  - **Statistical database** = set of data collected under the pledge of confidentiality for the purpose of producing statistics that do not compromise the privacy of the individuals that provided the data
- *Privacy goal*: Protect every individual while permitting statistical analysis of the DB as a whole. ("**Nothing is learned**" about an individual from before the query analysis to after).
- **Differential privacy process introduces randomness to queries in a structured way that produces the same statistical results.**

Differential Privacy Methodology

- Appropriately used, approach can be used to reasonably ensure privacy ***no matter what other data is available***

- Basic idea:
  - **Introduce randomness into the results of queries** to a statistical DB of confidential data. (Promotes privacy wrt "query release")
    - If done correctly, very accurate statistics of DB can be produced while ensuring high levels of privacy
  - Only a limited number of queries can be asked – there is a mathematically-determined **privacy budget**

- **Differential privacy is a condition of the release mechanism** (i.e. the trusted party releasing information about the dataset), rather than the dataset itself.

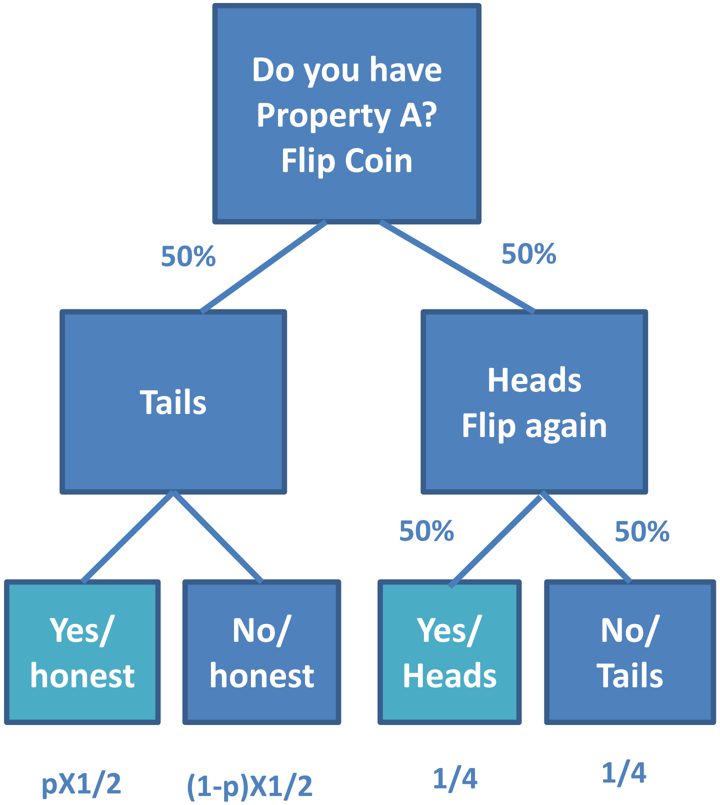
  - If two datasets are similar, a differentially private algorithm should behave roughly the same on both datasets (i.e. the presence or absence of an individual will not affect the final output of the algorithm significantly)

How does it work?

- **Procedure for achieving a randomized response** wrt query "Do you have Property A?"
  1. Throw a coin
  2. If tails, then answer honestly
  3. If heads, then throw the coin again and answer YES if heads, NO if tails

- **Solving for percentage of people with property A**
  - Let p = true proportion of people with Property A
  - By the procedure, we would expect to obtain pX1/2 + 1/4 = N YES answers.
  - If we know the number N of YES answers, we can figure out p without being able to figure out individual entries (p = 2N-1/2).

Simplified graphic, assumptions replacing Laplacian Distribution and other mathematical foundations:



Increasing use in private and public sectors

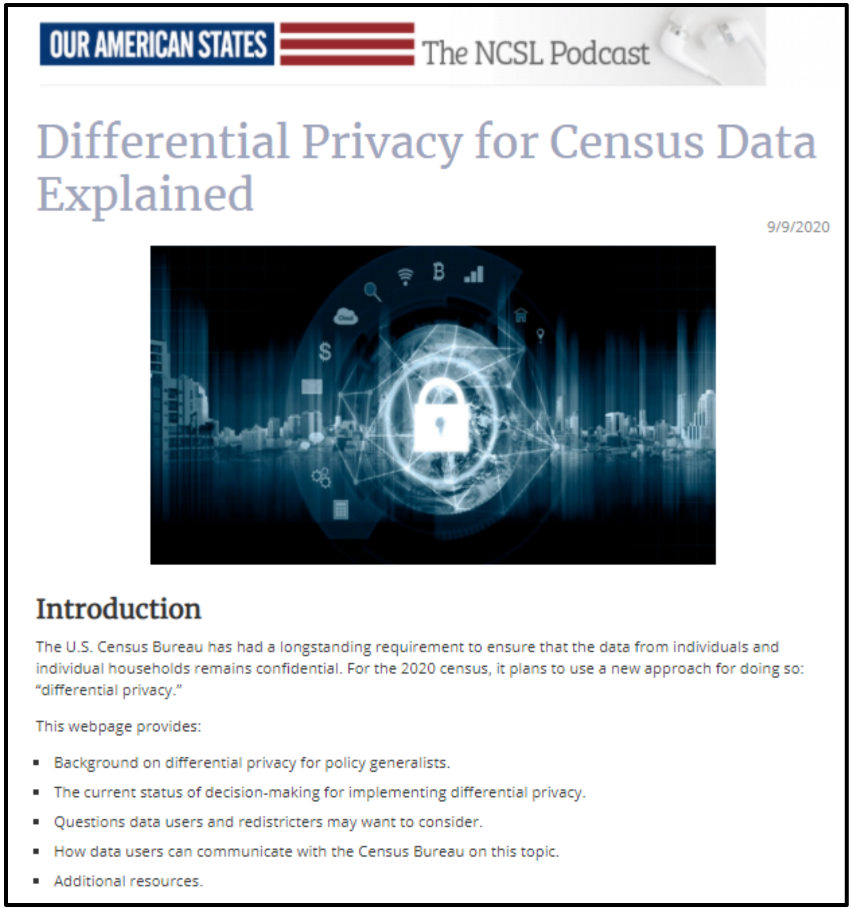
- **Adoption of differential privacy in real-world applications**
  - 2008: U.S. Census Bureau, for showing commuting patterns.
  - 2014: Google's RAPPOR, for telemetry such as learning statistics about unwanted software hijacking users' settings
  - 2015: Google, for sharing historical traffic statistics.
  - 2016: Apple announced its intention to use differential privacy in iOS 10 to improve its Intelligent personal assistant technology.
  - 2017: Microsoft, for telemetry in Windows.
  - 2019: Privitar Lens uses differential privacy API.
  - 2020: LinkedIn, for advertiser queries.

What are the limitations of differential privacy?

- Estimation from repeated queries -- with more and **more queries, privacy is breached**. (Exhausts the "privacy budget")
- **Data set must be "large enough"** so that statistical variations will work and promote accuracy
- Method works in situation where **queries about the group are of interest**, not in the situation where queries about the individual are important.
- Requires **substantial computation**
- **Trades-off privacy and accuracy**

Differential Privacy and the Census

- **“Differential Privacy for Census Data Explained”**, National Council of State Legislatures, https://www.ncsl.org/research/redistricting/differential-privacy-for-census-data-explained.aspx



Differential Privacy for Census Data Explained

9/9/2020

Introduction

The U.S. Census Bureau has had a longstanding requirement to ensure that the data from individuals and individual households remains confidential. For the 2020 census, it plans to use a new approach for doing so: “differential privacy.”

This webpage provides:

- Background on differential privacy for policy generalists.
- The current status of decision-making for implementing differential privacy.
- Questions data users and redistricters may want to consider.
- How data users can communicate with the Census Bureau on this topic.
- Additional resources.

U.S. Census Questions (for each household)

- The number of people living or staying in a home on April 1, 2020.
- Whether the home is owned with or without a mortgage or loan, rented or occupied without rent.
- A phone number for a “central” person in the home.
- For each person in the home:
  - Their name, sex, age, date of birth and race
  - Whether each person is of Hispanic, Latino or Spanish origin.
  - Where they usually live
  - The relationship of each person to a central person in the home.

What is Census data used for?

- Apportion seats in the House of Representatives
- Guide the allocation of approximately $675B in federal funds based on population counts, geography and demographic characteristics
- Support public and private sector decision making
- Serve as benchmark statistics for surveys and analyses throughout the subsequent decade
- Accuracy matters but Census Bureau works to minimize the likelihood that individuals can be reidentified in public data products

Privacy of Census data a bigger challenge than in the past

- Substantial information that can be coupled with Census data to reidentify individuals exists
- With publicly available 2010 Census and other information, Census researchers could re-identify 52 million people.
- Concern led to the use of differential privacy for 2020 Census
- Two ways to make data more private: limit the quantity of data released or reducing the accuracy of the data. Differential privacy provides a way to do some of both.

Implementing differential privacy for the Census tricky

- “Privacy loss budget” makes data accuracy and privacy competing uses of a finite resources.
  - How data is used matters and not all queries have the same cost to the privacy budget. Prioritization and evaluation of queries (e.g. for redistricting, public planning, demographic use, etc.) important.
- Formatting of data (e.g. ”moving” people between jurisdictions for postprocessing, population counts with fractional or negative values) can introduce bias and perturb results
- Users may need to focus on statistics, rather than data, creating a shift for the user community
- Census an important part of the larger discussion: what should be private and when is disclosure more important for the public good?

Lecture 7 References (not already on slides)

- **Differential Privacy**, Wikipedia
- **Differential Privacy, Simply Explained**, YouTube, https://www.youtube.com/watch?v=gI0wk1CXlsQ
- “**The Algorithmic Foundations of Differential Privacy**,” Cynthia Dwork and Aaron Roth, https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf
- ”**Differential Privacy: Seven lessons from the United States Census**”, Harvard Data Science Review, https://hdsr.mitpress.mit.edu/pub/dgg03vo6/release/2

Presentations

Upcoming Presentations

- Presentations for February 11
  - “**We’re banning facial recognition. We’re missing the point**.” New York Times, https://www.nytimes.com/2020/01/20/opinion/facial-recognition-ban-privacy.html
  - “**This site published every face from Parler’s Capitol riot videos**”, Wired, https://www.wired.com/story/faces-of-the-riot-capitol-insurrection-facial-recognition/
- February 18
  - “**Analysis: California privacy reboot puts rights in spotlight**”, Bloomberg Law, https://news.bloomberglaw.com/bloomberg-law-analysis/analysis-california-privacy-reboot-puts-rights-in-spotlight
  - “**To fix social media now, focus on privacy, not platforms**”, The Hill, https://thehill.com/opinion/technology/535824-to-fix-social-media-now-focus-on-privacy-not-platforms

Need Volunteers

February 22

- “**Grindr on the hook for 10M euro violations over GDPR consent violations**”, TechCrunch, https://techcrunch.com/2021/01/26/grindr-on-the-hook-for-e10m-over-gdpr-consent-violations/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAABSWJHh2KAv6uq4Ncppikhw3Ce8GDoEKFMOcJPFJy1kTbj1Fn_rpur6O7sq1LYNhqv1HzwQ7AVNLVUClRMg9wPBBVXTIxLK2WDqlMMtpFc68TjvPWzjrF0U4sqHCzns0wFJoubxi4WMlIoTy6bswMgd-YBJCxvHYwuGyB9scWgeT (Jeff H.)

- “**How the West got China’s social credit system wrong**,” Wired, https://www.wired.com/story/china-social-credit-score-system/ (Jin H.)

Presentations for Today

- Presentations for February 8
  - “**Changes to the census could make small towns disappear**,” New York Times, https://www.nytimes.com/interactive/2020/02/06/opinion/census-algorithm-privacy.html
  - “**Can a set of equations keep U.S. census data private?**,” Science, https://www.sciencemag.org/news/2019/01/can-set-equations-keep-us-census-data-private