DE-ID 201 Webinar Meeting Notes

(open to all education privacy working groups and Student Privacy Pledge signatories)

February 15, 2018
Topic: Differential Privacy
Hosts: Amelia Vance & Kelsey Finch, Future of Privacy Forum

<u>Moderator</u>: Michael Hawes, Direct of Student Privacy, US Dep't of Education

- Easiest to understand differential privacy if you put it in the context of where we've come from with privacy and public data.
- Since the 70's, our understanding of the risk of reidentification has expanded substantially—used to revolve around a few core identifiers.
- Now, that is recognized as inefficient.
- The toolbox also became more sophisticated but revolved around removing data, reducing precision of the data through aggregation or blurring, or perturbing the data to introduce noise or uncertainty.
    - But those techniques are largely more of an art form than a science in that they all serve to reduce the likelihood or ease with which a potential bad actor who's trying to find individuals, but it doesn't do so in any quantifiable kind of way.
- As computing technology has improved, and the amount of data about individuals and business has substantially increased over time, it has become more and more easy for those bad actors to perform that task.
- The discipline of differential privacy, which is less of a single technique and more of an approach, has its origins in the computer science world as a logical way of approaching privacy and anonymity from beta
- It was structured and developed to provide a way to give concrete guarantees about the privacy of and confidentiality of the data people are contributing in surveys, commercial settings, or in a census, in a variety of settings.
- Where differential privacy provides substantial promise to the data community is in its ability to quantity the risks and protections being afforded to data.
- A big concern that federal agencies often face is that you can apply protections today, but you don't know what risks against those same data in the future will be if more data become available.
- Differential privacy provides a scientific approach to protecting privacy; it's a promising technology.

<u>Differential Privacy with Applications to 2012 Census of Population</u>

John Abowd, Associate Director for Research and Methodology and Chief Scientist, U.S. Census

Why might database reconstruction matter of the U.S. Census Bureau

- We're having trouble controlling our ability to restrict reidentification in publications
- In 2003, a paper was published about the database reconstruction theorem that changed the privacy process in fundamental ways
    - The paper showed if you have a confidential database and you keep publishing statistics from it that are accurate, eventually, you expose the microdata you based it upon with near certainty.
- In order to design systems that protect statistical agencies from revealing too much information about their confidential databases in their own publications, they have to apply rules to them that

protect those same publications from any external data that someone might have as well into the indefinite future

2010 Census: 308,745,538 persons in the US population

- Also, the exact number of records in the database for which statistical tabulations were published
- High-level database schema (sample space)
    - Habitable blocks: 10,620,683
    - Habitable tracts:73,768
    - Sex: 2
    - Age values: 115
    - Race/Ethnicity (OMB Categories): 126
    - Race/Ethnicity (Summary File 2 Categories): 600
    - Relationship to person 1: 17
    - National histogram cells (OMB Categories): 492,660
- Summary of the publications (linearly independent)
    - Redistricting: around 2.8 billion statistics
    - Balance of Summary File 1: another 2.8 billion statistics
    - Summary File 2: 2.1 billion statistics
    - Public-use micro sample: 30.1 million statistics
    - Lower bound on published statistics: 7.7 billion statistics
    - Roughly 25 statistics published per person
- Reconstruction of the microdata is at least certainly a possibility when you're publishing that much of the microdata
- Bottom line: the database reconstruction theorem is the death knell for traditional data publication systems from confidential sources
    - Were not designed to protect against a reconstruction attack
    - Differential privacy doesn't protect against a reconstruction attack either, allows you to quantify the trade-off between data accuracy and privacy protection with reliable indicators

A little Economics

- Finite resource: information in an existing database
- Competing uses:
    - Accuracy (fitness-for-use) of published statistics
    - Loss of privacy (information leakage about individuals)
- Optimal resource allocation should equate:
    - Marginal rate of transformation (opportunity cost)
    - Marginal willingness to pay (marginal rate of substitution)
- Both accuracy and privacy are public goods
    - Private provision will generally produce sub-optimal accuracy and privacy loss
    - Non-rival public good: when you protect privacy with differential privacy, you aren't using up the privacy protection, one person's protection is at the expense of anyone else's
    - Title 13 § 9 prevents the Census Bureau from failing to protect data

Computer Science

- Formally private data protection systems (e.g., differential privacy, Dwork et al. 2006, but 2017 JPC is much clearer)
- Database definition, including neighboring databases
    - Have to set the privacy system up with formal definitions

        - Ask: what is the database, what is a neighboring database, what are you allowed to ask?
    - A neighboring database is one that looks just like the one you're trying to protect except something's been changed (one row's been dropped or the characteristics of 2 people in that database have been exchanged)
- Query sets
- Randomized query response mechanism
    - After figuring out what you are allowed to ask, you have to figure out how to add noise to these queries
        - This is the randomized query response mechanism, or the randomized publication mechanism
- Formal privacy definition
    - Differential privacy is a definition of if you're randomized query response mechanism can do this, then you afford this level of protection (usually called ε) to all of the people in all of the publications in the database, potentially or accuracy
- Measure of release data accuracy

Accuracy of Released Data

- Since randomized query response mechanisms add noise to the correct answer from the database, an accuracy measure is required to compare the protected answer to the true answer
- There are many accuracy measures that might be suitable
    - Depends on what application you are trying to make suitable for use
- Most depend on the absolute difference between the true answer and the released answer
- We will confine attention today to two accuracy measures that have this property:
    - Normalized total variation distance ($1–0.5L_1$ distance/N)
    - Statistical precision relative to the precision in the confidential data (function of L_2 distance)
- You don't talk about the actual accuracy, you talk about the expected accuracy

Example: Differential Privacy for Simple Tables

- The simple tables in this example are based on the 2010 Census of Population PL94-171 release, known popularly as the redistricting data
- These data are used to redraw every legislative district from Congressional all the way down to village councils in every state every 10 years
- They must be released by April 1st of the year following a decennial census
    - for the 2010 Census, April 11, 2011
- Census block - risk involved [*geographic zoom-in of just outside Ithaca, NY called Cayuga Heights is shown; each purple line shows a district block*]
    - I often know exactly who is inside of these blocks because they are my colleagues at Cornell, which is why we often say that disclosing data at a block level is at a high risk of re-identification
- Levels of geography for which Bureau publishes redistricting data
    - States: 51
    - Counties: 3,143 or equivalents
    - Tracts: 73,768; excludes water only
    - Block Groups: 220,135; excludes water only
    - Blocks: 10,620,683; excludes water only
        - All of those OMB race blocks are applied at a block level

Disclosure Limitation is a Technology

- The price of increasing data quality (public good) in terms of increased privacy loss (public bad) is the slope of the technology frontier:
  - Economics: production possibilities frontier (risk-return in finance)
  - Forecasting models: receiver operating characteristics curve
  - Statistical disclosures limitation: risk-utility curve (with risk on the x-axis)
- All the exact same thing—describe a technology, don't tell you how to pick a point on it

How Do You Set ε?

- Dwork (2008): The parameter ε in Definition 1 is public. The choice of ε is essentially a social question and is beyond the scope of this paper.
- Dwork (2011): The parameter ε is public, and its selection is a social question. We tend to think of ε as, say, .01, .1, or in some cases, In 2 or In 3.
- In OnTheMap, the Census Bureau set ε to equal 8.9
  - Required to produce tract-level estimates with acceptable accuracy
- Apple claim: ε= 2.4, or 8 according to the application
  - It's been reversed engineered, and those that have done so claim it's really closer to 14 (Tang et al.)
- Google claim (Chrome Browser): ε = 4.39
  - It's been reversed engineered, and those that have done so claim it's really closer to 9 (McSherry)
- All these settings are differentially private, but they have very different global disclosure risks
- Data stewardship policy committee sets ε

Choice Problem for Redistricting is More Challenging

- In the redistricting application, the fitness-for-use is based on:
  - Supreme Court one-person one-vote decision
    - Congressional districts have a tolerance of +/- person; state legislatures +/- 5%; some states are stricter
    - Some states stricter, California: +/- 5 people
  - Voting Rights Act, § 2: Requires majority minority districts at all levels, when certain criteria
  - The accuracy has to allow these two things to happen
- The privacy interest is based on:
  - Title 13 requirement not to publish exact identifying information
  - The public policy implications of other uses of detailed population race and ethnicity.
- At Bureau we've been using a variety of technologies to minimize re-identification risk:
  - Randomized response: most efficient input noise infusion mechanism
  - Parallel-composed geometric mechanism: most efficient output noise infusion mechanism for a workload with sensitivity one.
  - Customized query selection or workload management (e.g., matrix mechanism): more efficient output noise infusion mechanism for correlated queries

Takeaways

- This is new ground for official statisticians, we need to start talking about the social choice problem

    - They do not have to think explicitly about the social choice problem of competing interests on accuracy and privacy using traditional disclosure avoidance methods
    - But, it is not foreign territory, since the invention of sampling theory, official statisticians have thought about the tradeoff between accuracy and all aspects of the design of surveys

Differential Privacy: A Primer for a Non-technical Audience

Alexandra Wood, Research Fellow and Senior Researcher for the Privacy Tools Project, Harvard's Berkman Klein Center for Internet & Society

Data Privacy: The Problem

- When speaking about differential privacy, we're talking about privacy in the context of a statistical computation
    - This refers to any analysis or computation that takes personal data and transforms it into some outputs
    - This can be thought of broadly as outputs used for scientific inquiry, policy-making decisions, investment decisions, etc.
- Central question: how can personal data be analyzed and shared, while ensuring the privacy of the individuals in the data will be protected?

Real-World Example of Privacy Attack

- The late 1990s, Massachusetts Group Insurance Commission allowed research groups to access anonymized records with information about all hospital visits made by state employees
- Before doing so, the agency removed names, dates of birth, social security numbers, and other pieces of information that could be used to identify individuals in the data
- Professor Latanya Sweeney, MIT, set out to identify one of the records, choosing the Governor of Massachusetts William Weld
    - Obtained information about him: zip code, birth date, gender, through voter registration records
    - Finding just one record that matched these three data points allowed her to mail the Governor a copy of his personal medical records
- This case illustrates a point that although a data point may appear to be anonymous, it may be used along with other data to identify individuals
- A series of attacks have similarly been carried out to re-identify individuals, illustrating that the risk remains even where additional pieces of information are removed

Attacks on Privacy: Key Takeaways

- Lack of rigor leads to unanticipated privacy failures
    - New attack modes emerge as research progresses
    - Redaction of identifiers, release of aggregates, etc. is insufficient to protect privacy
    - Auxiliary information must be taken into consideration
- Any useful analysis of personal data must leak some information about individuals
- Information leakages accumulate with multiple analyses/releases
- Mathematical facts, not matters of policy

Emergence of Differential Privacy

- A new line of privacy work in theoretical computer science (beginning ~2003)
- Yields a new concept: differential privacy (2006)

    - Supported by rich theory
    - In its first stages of implementation and real-world usage
        - US Census, Google, Apple, Uber, etc.

What is Differential Privacy?

- Differential privacy is a definition (or standard) of privacy
    - Not a specific technique or algorithm
- It expresses a specific desiderata of analysis:
    - Any information-related risk to a person should not change significantly as a result of that person's information being included, or not, in the analysis

A Privacy Desiderata

- Consider a scenario, such as estimating the number of people on this webinar call who have red hair
- Ideally, this estimate should remain exactly the same, whether or not an individual such as myself is included in this survey. However, insuring this property exactly would require the total exclusion of my information from the computation
    - If I remove my information, and we want to protect other people from the call, and continue with this line of argument for everyone on the call, we'll come to the conclusion that everyone's information must be removed in order to satisfy that person's ideal scenario

A More Realistic Privacy Desiderata

- To overcome this dilemma, differential privacy requires only that the output of the analysis remain approximately the same, whether I participate in the survey or not
- Allows for deviation between real-world setting and ideal-world setting
    - Privacy parameter ε quantifies and limits the deviation between these two scenarios
    - Referred to as the *privacy loss parameter*

Understanding Differential Privacy

- Because differential privacy is a standard, it can be interpreted in different ways
    - Looked for comparisons to explain the relationship from other disciplines, such as the law
        - Ex: some laws require that individuals be given the opportunity to "opt out" of a data release—differential privacy can be viewed as an automatic opt-out
- Provides an "automatic opt-out:" Differential privacy essentially protects an individual's information as if her information were not used in the analysis at all.
- Protects personally identifiable information: differential privacy essentially ensures that using an individual's data will not reveal any personally identifiable information that is specific to her.
    - Here, *specific* refers to information that cannot be inferred unless the individual's information is used in the analysis.
- Protects against inferences: differential privacy essentially masks the contribution of any single individual, making it impossible to infer any information specific to an individual, including whether the individual's information was used at all
- Differential privacy provides protection (far) beyond notions of "identifiability"

Michael Hawes: Just to clarify, these guarantees, like with the absolute statements here would be if ε was set zero, these would be absolute, but with ε as a number above zero that the guarantee up to a certain level of certainty, correct?
Alexandra Wood: That's correct, and that's what the qualifying here essentially is referring to is ε.

Achieving Differential Privacy

- Essential component of differential privacy computation is the privacy loss of parameter ε of how much noise is added to a computation, and can be thought of as a privacy tuning knob
- Algorithms maintain differential privacy via the introduction of carefully crafted random noise into the computation
- Increases in ε result in less noisy computations, but also less privacy protection

Combining Differentially Private Analyses

- Combinations of ε-differentially private computations are also differentially private (with larger ε)
- This is important for protecting privacy, as every analysis results in some leakage of information about the individuals whose information is being analyzed and that this leakage accumulates with each analysis
- It is also a (unique) feature of differential privacy
  - Most, if not all known definitions of privacy do not measure the cumulative risk for multiple analyses or release of data about the same individuals

Real-World Implementations

- US Census Bureau
- Google (Chrome; Malware)
- Apple (usage statistics through iOS 10)
- Uber (average length of ride, other statistics on usage)

Conclusion

- Provides protection that is robust to a wide range of potential privacy attacks, including attacks unknown at the time of deployment
- Protects privacy independent of the methods and resources used by a potential attacker
- Provides provable privacy guarantees with respect to the cumulate risk of successive data releases
- Has the benefit of transparency, as it is not necessary to maintain secrecy around a differentially private computation or its parameters
- Can be used to provide broad public access to data in a privacy-preserving way

Michael Hawes: What types of information does differential privacy actually protect?
Alexandra Wood: Differential privacy protects any information that's specific to an individual. Unlike traditional techniques that focus on suppressing specific pieces of information, differential privacy doesn't distinguish between identifying and non-identifying information and provides protection for all information that could be learned about an individual based on that individual's participation in a data analysis.

Michael Hawes: What types of analyses that can be performed, any types that can't be calculated based on a differentially private approach?
Alexandra Wood: A large number of different types of statistical analyses can be performed with differential privacy guarantees, for example there are algorithms known to exist for producing simple

counts, histograms, cumulative distribution functions, linear regressions, and machine learning analyses among many others with differential privacy guarantees. It can even be used to produce synthetic microdata or individual-level data like what John described with the Census Bureau application.

Michael Hawes: To interpret that to some degree, the data that come out of differentially private analyses are generally either calculations that've had noise added to them to move the number slightly or they're actually synthetic data generated from the underlying patterns in the original broad data, correct?
Alexandra Wood: That's correct, yes.

Audience question: Are you aware of any commercial solutions that allow applying differential privacy to data sets?
Alexandra Wood: I've heard that there are companies developing tools like that, I'm not sure any are available yet.

Following up on that, if an organization were interested in using differential privacy for their public data releases, what kind of burden would that entail, what skills would they need to have in order to start applying it?
Alexandra Wood: In the near term, having a background in computer science is a prerequisite. There are some pieces of source code on GitHub that have been made available as prototypes for researchers to work with, but I think moving forward what's needed is off-the-shelf tools that people with a background in differential privacy could use. Members of our team have been working to develop differential privacy tools, through the Harvard data-verse, that lay-audiences could use.

Audience question: Can you speak about the effect of sample size or database size on the usability of the resulting data for any given privacy budget or ε value?
Alexandra Wood: Generally, the longer the data set, the more records there are, the more accurate your statistics can be when using differential privacy for a given level of privacy protection, though high-dimensional data, wider data with more attributes per record, can be more challenging. In those cases, interactive mechanisms are often chosen, such as query-based approaches where data users can submit queries that allow the computation of statistics on certain attributes at a time.

Audience question: With any differentially private implementation short of going the synthetic data route, each analysis that you perform counts against your privacy budget, your ε there, what happens when you hit the analysis that actually meets or exceeds your ε, what happens to future queries after that?
Alexandra Wood: Where there are a couple of approaches. You could have a budget set for each user, based on the assumption that users aren't colluding and comparing answers. Another solution is synthetic data, which only has an effect on the privacy budget once.

Audience question: Are there any potential attacks against differential privacy that are of concern?
Alexandra Wood: I'm not the best person to answer the question, I'm not as familiar with attacks. I understand that there may be some side-channel attacks.

Audience questions: Are there any good resources for getting started on actually implementing a differentially private approach to data?

Alexandra Wood: I’ve discussed a number of different implementations by different organizations as well last research at Harvard and other research institutions to develop general-purpose tools of differential privacy.