Data Security and Privacy

Topic 22: Meaning and Caveats of Differential Privacy

Semantic Interpretation of DP?

- Impossibility result means that bounding difference from prior and posterior is hard
- Approach 1: Provide posterior-to-posterior bound
  - Identify “ideal worlds” where individuals’ privacy are preserved: the i’th ideal world is to remove the i’th individual’s data
  - Bound differences between ``ideal worlds’’ and the “real world”
- Approach 2: Understand under which condition prior-to-posterior bound can be ensured

Our Formulation of DP's Real-World Ideal-World Privacy Guarantee

- Adversary is modeled as a decision function f, which after observing a transcript t, makes a decision among C.
- The prob an adversary chooses c after interacting with *A(D)* is
  - $Adv^{A(D)}(c) = \sum_t \Pr[A(D) = t] * f(t)[c]$
- *Def:* α-opting-out-simulation: Let D' be the dataset resulted from an individual opting out from D, then
  - $e^{-\alpha} Adv^{A(D')}(c) \leq Adv^{A(D)}(c) \leq e^{\alpha} Adv^{A(D')}(c)$
- Thm: ε-DP is equivalent to α-opting-out-simulation

DP’s Similar-Decision-Regardless-of-Prior Guarantee

- Regardless of external knowledge, an adversary with access to the sanitized database makes similar decisions whether or not one individual’s data is included in the original database.

The Personal Data Principle

- Data privacy means giving an individual control over his or her personal data. An individual's privacy is not violated if no personal data about the individual is used.
- Privacy does not mean that no information about the individual is learned, or no harm is done to an individual; enforcing the latter is infeasible and unreasonable.

OECD Privacy Principles

- **1. Collection Limitation Principle**
  - There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.
- **2. Data Quality Principle**
  - Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.

OECD Privacy Principles

- **3. Purpose Specification Principle**
  - The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.
- **4. Use Limitation Principle**
  - Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with Principle 3 except:
  - a) with the consent of the data subject; or
  - b) by the authority of law.

OECD Privacy Principles

- **5. Security Safeguards Principle**
  - Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data.
- **6. Openness Principle**
  - There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

OECD Privacy Principles

- **7. Individual Participation Principle**
  - An individual should have the right:
  - a) to request to know whether or not the data controller has data relating to him;
  - b) to request data relating to him, …
  - c) to be given reasons if a request is denied; and
  - d) to request the data to be rectified, completed or amended.
- **8. Accountability Principle**
  - A data controller should be accountable for complying with measures which give effect to the principles stated above.

Genius of Idea Behind DP

- Privacy is hard, because information may correlate
- By identifying a world without one individual’s data as an ideal world for the individual, DP does not need to deal with data correlation
- This insight is summarized by the personal data principle

Critique of DP

- From [Kifer and Machanavajjhala, 2011]
- *“Additional popularized claims have been made about the privacy guarantees of differential privacy. These include:*
  - *It makes no assumptions about how data are generated.*
  - *It protects an individual’s information (even) if an attacker knows about all other individuals in the data.*
  - *It is robust to arbitrary background knowledge.”*

Kifer and Machanavajjhala: No Free Lunch in Data Privacy, SIGMOD 2011.

An Attempt at Providing Prior-to-Posterior Bound in [Dwork et al. 2006]

- A mechanism is said to be **(k, ε)-simulatable** if for **every informed adversary** who already knows all except for k entries in the dataset D, every output, and every predicate f, the change in the adversary's belief on f is multiplicative-bounded by e^{ε}.
- Thm: ε-DP is equivalent to (1,ε)-simulatable.
- Does this mean ε-DP provides prior-to-posterior bound for an arbitrary adversary?
  - Wouldn't that conflict with the impossibility results?

Dwork et al.: Calibrating Noise to Sensitivity in Private Data Analysis. TCC 2006.

An Example Adapted from [Kifer and Machanavajjhala, 2011]

- Bob or one of his 9 immediate family members may have contracted a highly contagious disease, in which case the entire family would have been infected. An adversary asks the query “how many people at Bob's family address have this disease?”
- What can be learned from an answer produced while satisfying ε-DP?
  - Answer: Adversary’s belief change on Bob’s disease status may change by something close to $e^{10\varepsilon}$.
- Anything wrong here?

In A Sense, No

1. An adversary’s belief about Bob’s disease status may change by a factor of $e^{10\epsilon}$ due to data correlation. This is an example that DP cannot bound prior-to-posterior belief change against arbitrary external knowledge.
2. DP’s guarantee about posterior-to-posterior bound remains valid.
3. The analysis in [Dwork et al. 2006] is potentially misleading, because it could lead one to think that DP can offer more protection than it actually does.
   – The notion of informed adversary, while appearing strong, is in fact, very limiting.
4. Applying PDP, ε-DP is doing what it is supposed to do, but stay tuned

Caveats of Applying DP

- How neighboring datasets is defined?
- Whether composition is considered in the local setting
- What constitutes an individual’s data
- One individual’s data or personal data under one individual’s control
- Group privacy
- Moral challenge
- Choosing epsilon value
- Learning models and applying to individuals
- Privacy and discrimination

Defining Neighbors Incorrectly

- Edge-DP in graph data is inappropriate
  - Typically one individual controls a node and its relationship.
  - ``Attacks’’ on graph anonymization typically in the form of node identification.
  - Suppose the goal is to protect edge info, then edge-DP still fails, because of correlation between edges.
- Packet-level privacy for networking data is inappropriate
- Cell-level privacy in matrix data is usually inappropriate

Local Setting

- Google's RAPPOR system is not good enough
  - Erlingsson et al. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. CCS 2014.
  - One system may collect answers to many questions; and each question is answered with privacy budget ε
- Apple seems to be doing the same

What Constitutes An Individual’s Personal Data?

- Is the genome of my parents, children, sibling, cousins “my personal information”?

- Example: DeCode Genetics, based in Reykjavík, says it has collected full DNA sequences on 10,000 individuals. And because people on the island are closely related, DeCode says it can now also extrapolate to accurately guess the DNA makeup of nearly all other 320,000 citizens of that country, including those who never participated in its studies.

Such legal and ethical questions still need to be resolved

- Evidences suggest that such privacy concerns will be recognized.
- In 2003, the supreme court of Iceland ruled that a daughter has the right to prohibit the transfer of her deceased father's health information to a Health Sector Database, not because her right acting as a substitute of her deceased father, but in the recognition that she might, on the basis of her right to protection of privacy, have an interest in preventing the transfer of health data concerning her father into the database, as information could be inferred from such data relating to the hereditary characteristics of her father which might also apply to herself.

https://epic.org/privacy/genetic/iceland_decision.pdf

Lesson

- When dealing with genomic and health data, one cannot simply say correlation doesn't matter because of Personal Data Principle, and may have to quantify and deal with such correlation.

My Personal Data or Personal Data Under My Control?

- Consider the following variants of the Bob example.
- Case (a). Bob lives in a dorm building with 9 other unrelated individuals. Either they all have the disease or none. One can query how many individuals at this address have the disease.
- Case (b). The original example: Bob and 9 family members.
- Case (c). Bob and 9 minors for which Bob is the legal guardian.

Our Tentative Answer

- Case (a). Bob and 9 other unrelated individuals.
  - DP does what it suppose to do based on Personal Data Principle.
- Case (b). The original example: Bob and 9 family members.
  - Difficult to say: on the borderline and not enough information.
- Case (c). Bob and 9 minors
  - Using DP this way is inappropriate, because Bob controls the 9 other records as well, and

Group Privacy as a Potential Challenge to Personal Data Principle

- Can a group of individuals, none of whom has specifically authorized usage of their personal information, together sue on privacy grounds that aggregate information about them is leaked?
  - If so, satisfying DP is not sufficient.
  - Would size of group matter?

A Moral Challenge to DP

- Question from Quora:
  - Say I steal 2 cents from every bank account in America. I am proven guilty, but everyone I stole from says they're fine with it. What happens?

- If one makes profit from applying DP to a dataset of many individuals, isn’t this morally the same as the above?

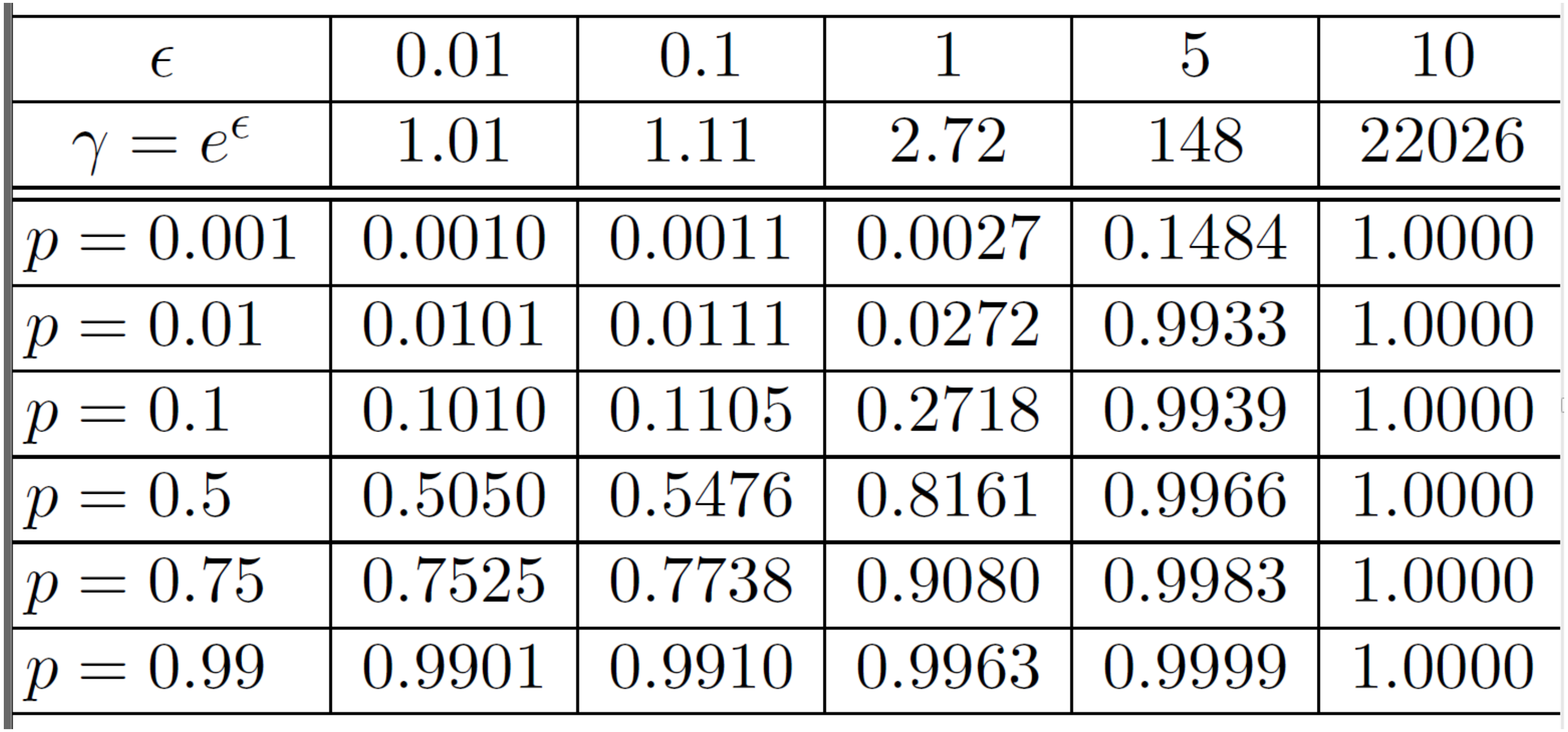
How to Choose ε

- From the inventors of DP: "*The choice of* ϵ *is essentially a social question. We tend to think of* ϵ *as, say,* 0.01*,* 0.1*, or in some cases,* $\ln 2$ *or* $\ln 3$".

- Our position.
  - ϵ of between 0.1 and 1 is often acceptable
  - ϵ close to 5 might be applicable in rare cases, but needs careful analysis
  - ϵ above 10 means very little

- Why?

Consult This Table of Change in Belief: p is prior; numbers in table are posterior



ϵ	0.01	0.1	1	5	10
$\gamma = e^{\epsilon}$	1.01	1.11	2.72	148	22026
$p = 0.001$	0.0010	0.0011	0.0027	0.1484	1.0000
$p = 0.01$	0.0101	0.0111	0.0272	0.9933	1.0000
$p = 0.1$	0.1010	0.1105	0.2718	0.9939	1.0000
$p = 0.5$	0.5050	0.5476	0.8161	0.9966	1.0000
$p = 0.75$	0.7525	0.7738	0.9080	0.9983	1.0000
$p = 0.99$	0.9901	0.9910	0.9963	0.9999	1.0000

Apply a Model Learned with DP Arbitrarily.

- There are two steps in Big Data
  - Learning a model from data from individuals in A
  - Apply the model to individuals in B, using some (typically less sensitive) personal info of each individual, one can learn (typically more sensitive) personal info.
    - The sets A and B may overlap
- The notion of DP deals with only the first step.
- Even if a model is learned while satisfying DP, applying it may still result in privacy concern, because it uses each individual’s personal info.

The Target Pregnancy Prediction Example

- Target assigns every customer a Guest ID number and stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources.
- Looking at historical buying data for all the ladies who had signed up for Target baby registries in the past, Target's algorithm was able to identify about 25 products that, when analyzed together, allowed Target to assign each shopper a ``pregnancy prediction'' score.
- Target could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

Privacy and Discrimination

- What if one applies a classifier to public information (such as gender, age, race, nationality, etc.) and make decisions accordingly
- Is there privacy concern?
- Better privacy may cause more discrimination!
  - From Wheelan's book "Naked Economics"
  - Hiring blacks with (and w/o) criminal background checks.

When is ϵ-DP Good Enough?

- Applying ϵ-DP in a particular setting provides sufficient privacy guarantee when the following conditions hold:
  - (0) Group privacy / morality challenges do not hold
  - (1) The Personal Data Principle can be applied;
  - (2) All data one individual controls are included in the difference of two neighboring datasets;
    - With (1) and (2), even if some information about an individual is learned because of correlation, one can defend DP.
  - (3) An appropriate ϵ value is used.

Next Lecture

- Revising Local DP