Differential Privacy: Potential and Limitations

Ninghui Li

Department of Computer Science
Purdue University

Defining Privacy is Hard

- Lots of privacy notions
  - E.g., k anonymity, l diversity, t closeness, and many others
- Why defining privacy is hard?
  - Difficult to agree on adversary goal.
  - Difficult to agree on adversary power.
  - Too strong , then not achievable.
  - Too weak, then not enough
  - Information is correlated.

What is Privacy?

It is complicated!

Some concepts from the book “Understanding Privacy” by Daniel J. Solove:

1. the right to be let alone
2. limited access to the self
3. secrecy—the concealment of certain matters from others;
4. control over others' use of information about oneself
5. personhood—the protection of one’s personality, individuality, and dignity;
6. intimacy—control over, or limited access to, one’s intimate relationships or aspects of life.

Impossibility of "Privacy as Secrecy"

- Dalenius [in 1977] proposes this as privacy notion: "*Access to a statistical database should not enable one to learn anything about an individual that could not be learned without access*."
  - Similar to the notion of semantic security for encryption
  - Requires a **prior-to-posterier** bound
  - Not possible if one wants utility.

An Example of the Impossibility of Providing Prior-Posterior Bound

- Assume that smoking causes lung cancer is not yet public knowledge, and an organization conducted a study that demonstrates this connection and now wants to publish the results.
- A smoker Carl was not involved in the study, but complains that publishing the result of this study affects his privacy, because others would know that he has a higher chance of getting lung cancer, and as a result he may suffer damages, e.g., his health insurance premium may increase.

- Can Carl legitimately complain about his privacy being violated by publishing results of the study?

Analogies with Crypto

- Semantic security can be achieved. Why can't we achieve a privacy notion similar to semantic security?
  - There are two kinds of recipients in encryption, but only one in the setting for privacy.
- What about order/property-preserving encryption?
  - Security defined as simulating an ideal world
- "Real-world-ideal-world" approach also used in Secure Multiparty Computation

Differential Privacy [Dwork et al. 2006]

- Definition: A mechanism *A* satisfies ε-Differential Privacy if and only if
  - for any neighboring datasets D and D'
  - and any possible transcript $t \in$ Range(A),
$$\Pr[A(D) = t] \leq e^{\epsilon} \Pr[A(D') = t]$$
  - For relational datasets, typically, datasets are said to be neighboring if they differ by a single record.
- Intuition:
  - Output does not overly depend on any single record
  - Privacy is not violated if one's information is not included in the input dataset

Genius of Idea Behind DP

- Privacy is hard because which information to hide is difficult to enumerate and information may correlate
- By identifying a world without one individual's data as an ideal world for the individual, and providing real-world-ideal-world bound, one does not need to provide prior-to-posterior bound, and does not need to deal with data correlation
- DP simulates privacy is "control over others' use of information about oneself"

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.

Caveats of Applying DP

- How neighboring datasets is defined
- How many pieces of information from one user is collected in the local setting
- What constitutes an individual's data
- Group privacy
- Moral challenge
- Choosing epsilon value
- Learning models and applying to individuals

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 DP for networking data is inappropriate
- Cell-level DP in matrix data is usually inappropriate
- Pixel-level DP for image is meaningless
- Single-picture level DP where one individual has many pictures is likely 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.

An Example Adapted from [Kifer and Machanavajjhala, 2011]

- 10 people live together (e.g., in a fraternity house). They may have contracted a highly contagious disease, in which case all would have been infected. An adversary asks the query "how many people at this 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?

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.

https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

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.

Thank You

- Questions?