CS208: Applied Privacy for Data Science
Beyond Noise Addition & Tabular Data

School of Engineering & Applied Sciences
Harvard University

February 21, 2022

Announcements

- Salil’s OH this week: today 2-3pm Zoom, Thu 1-2pm in person
- Please keep using name placards. Thu: introductions in class.
- Fill out midterm feedback survey by tomorrow (Wed) eve.
- Final Project ideas on HW4
  - This is initial, solo brainstorming
  - Look at annotated course bibliography for literature pointers
  - Revisions after spring break, possibly with partners
- Privacy Tools/OpenDP opportunities
  - Summer internships: email privacytools-info@seas.harvard.edu
  - Data Science/CSE Masters’ theses

Beyond Noise Addition

Fact: Laplace (or geometric) mechanism has the optimal worst-case, additive accuracy for answering a single real-valued query with DP.

Q: Why might we want other mechanisms?

Approaches we'll see:

1. Decompose a computation into several queries that can be answered with Laplace (cf. HW4 #2)
2. The exponential mechanism: applies even to discrete outcomes (ex: median)
3. Approaches based on local sensitivity and restricted sensitivity (ex: graph statistics → beyond tabular data)

DP Medians

Recall: median over $\mathcal{X} = [0,1]$ has global sensitivity $\geq 1/2$

- Laplace Mechanism is useless.
- Same applies for discrete data, e.g. education variable $\in \{\mathrm{bb}, 1,2, \ldots, 16\}$ in PUMS dataset.

Q: what mechanism from prior weeks could use to estimate the median of a discrete variable?

Exponential Mechanism

Given output space $\mathcal{Y}$ and score function $s\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the exponential mechanism is:

$$M(x)\text{: output } y \in \mathcal{Y} \text{ with probability} \propto \exp\left(\varepsilon \cdot \frac{s(x,y)}{2 \cdot \mathrm{GS}_s}\right),$$

$$\text{where } \mathrm{GS}_s \stackrel{\text{def}}{=} \max_{x \sim x', y} |s(x,y) - s(x',y)|.$$

Thm: above mechanism is ε-DP.

Q: what is a good score function for the median?

Utility of the Exp Mech Median

See notebook for experiments comparing using exponential mechanism vs. histograms for the median.

Theoretical explanation: whp

- Exp mech outputs y with $q(x, y) \geq n/2 - O(\log|\mathcal{Y}|)/\varepsilon$.
- Histogram outputs y with $q(x, y) \geq n/2 - O(|\mathcal{Y}|^{1/2})/\varepsilon$.

Q: How to implement for continuous $\mathcal{Y} = [0,1]$?

Local Sensitivity

- Even when GS_q is large, the local sensitivity may often be small on many natural datasets x:
$$\mathrm{LS}_q(x) \overset{\text{def}}{=} \max_{x' : x' \sim x} |q(x') - q(x)|.$$

- Adding noise proportional to local sensitivity is not DP (why?)

- But there are several DP methods that approximate this idea (smooth sensitivity, propose-test-release, privately bounding local sensitivity, restricted sensitivity).

- We'll illustrate a couple of these approaches on graph & network data.

DP for Graph and Network Data

- Datasets: graphs G, possibly with labels on nodes & edges

- Edge-level privacy: $G \sim G'$ if differ on at most one edge
  - Only protects particular relationships

- Node-level privacy: $G \sim G'$ if differ on at most one node and incident edges
  - "Right" definition to protect individual info in a social network

Publishing information about graphs

Many types of data can be represented as graphs where

- nodes correspond to individuals
- edges capture relationships
  - “Friendships” in online social network
  - Financial transactions
  - Email communication
  - Health networks (of doctors and patients)
  - Romantic relationships

image source http://community.expressor-software.com/blogs/mtarallo/36-extracting-data-facebook-social-graph-expressor-tutorial.html

facebook
Search
invite friends | click anywhere to drag | mouse wheel to zoom at mouse pointer | double-click nodes to visit profile pages
zoom

image source http://www.queticointernetmarketing.com/new-amazing-facebook-photo-mapper/

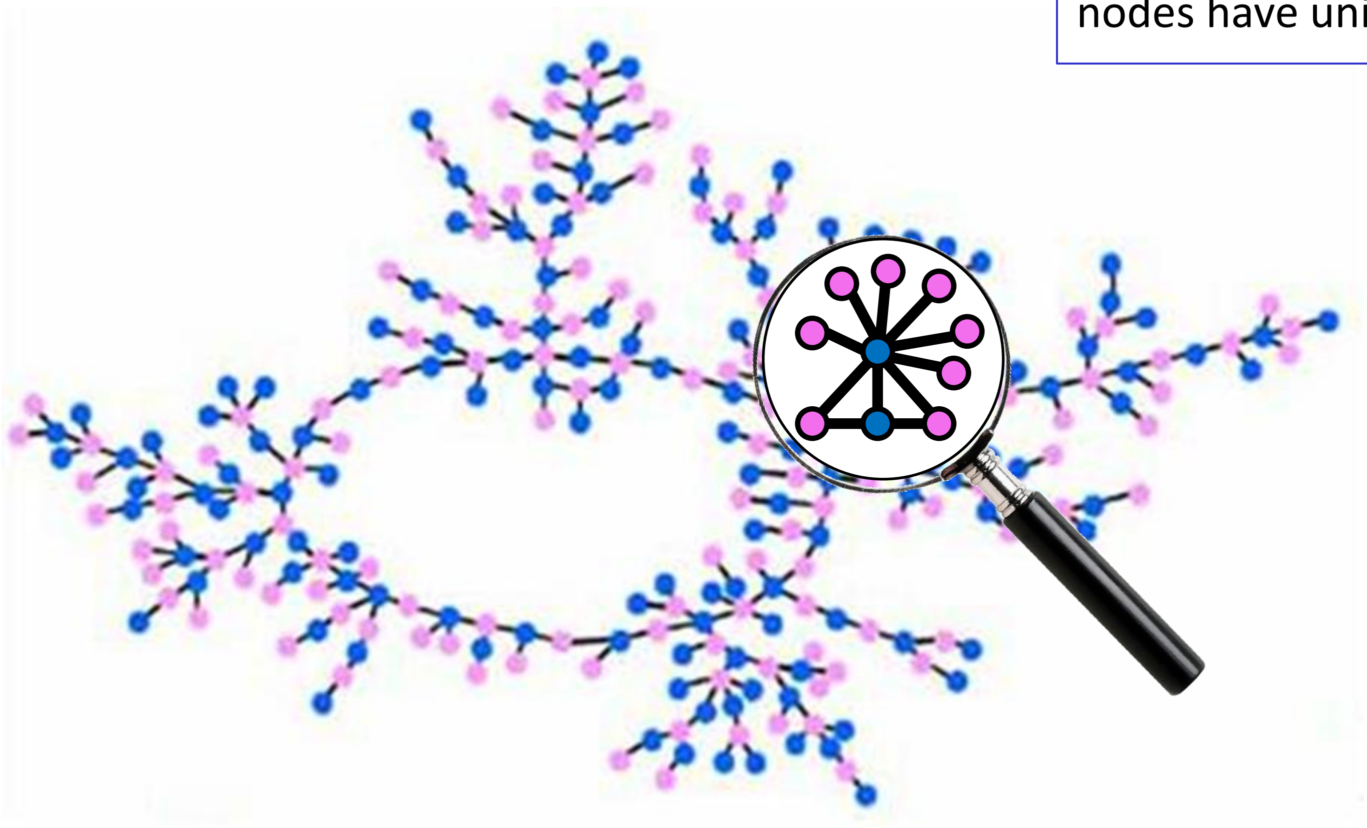
Such graphs contain potentially sensitive information.

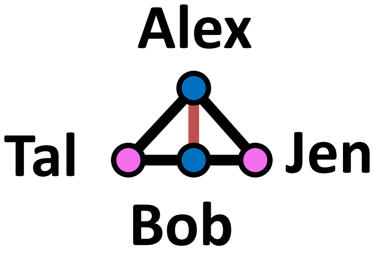
[slide from Sofya Raskhodnikova]

“Anonymized” graphs still pose privacy risk

- False dichotomy: personally identifying vs. non-personally identifying information.
- Links and any other information about individual can be used for de-anonymization.

In a typical real-life network, many nodes have unique neighborhoods.





Bearman, Moody, Stovel. Chains of affection: The structure of adolescent romantic and sexual networks, American J. Sociology, *2008*

[slide from Sofya Raskhodnikova]

Sensitivity of Graph Queries

query	range	$\mathbf{LS}_q^{\text{edge}}(G)$	$\mathbf{GS}_q^{\text{edge}}$	$\mathbf{LS}_q^{\text{node}}(G)$	$\mathbf{GS}_q^{\text{node}}$
# edges	$[0, dn/2]$				
max-deg	$[0, d]$				
# triangles	$[0, \Theta(nd^2)]$				

$d = \mathrm{maxdeg}(G)$, typically $d \ll n$

Sensitivity of Graph Queries

query	range	$\mathbf{LS}_q^{\text{edge}}(G)$	$\mathbf{GS}_q^{\text{edge}}$	$\mathbf{LS}_q^{\text{node}}(G)$	$\mathbf{GS}_q^{\text{node}}$
# edges	$[0, dn/2]$				
max-deg	$[0, d]$				
# triangles	$[0, \Theta(nd^2)]$				

$d = \text{maxdeg}(G)$, typically $d \ll n$

A $(2\varepsilon, \delta)$-DP mechanism $M(G)$ for # triangles under edge privacy:

1. Let $\hat{d} = \text{maxdeg}(G) + \text{Lap}\left(\frac{1}{\varepsilon}\right) + \frac{\ln(1/\delta)}{\varepsilon}$.
2. Output $\#\text{triangles}(G) + \text{Lap}\left(\frac{\hat{d}}{\varepsilon}\right)$.

Why does this work?

Restricted Sensitivity

[Blocki-Blum-Datta-Sheffet `13]

Let:

- $\mathcal{G}$ = set of all datasets, with neighbor relation $\sim$.
- $d(x, x')$ = distance between datasets $x, x' \in \mathcal{G}$ wrt $\sim$
  = smallest k s.t. ∃ path $x = x_0 \sim x_1 \sim \cdots \sim x_k = x'$.
- $\mathcal{H} \subseteq \mathcal{G}$ be a "hypothesis" set of datasets in which we hope that our dataset lies.

Def: For $q : \mathcal{G} \to \mathbb{R}$, the restricted sensitivity of q wrt $\mathcal{H}$ is:

$$\mathrm{RS}_q^{\sim,\mathcal{H}} = \max_{x,x' \in \mathcal{H}} \frac{|q(x) - q(x')|}{d(x, x')}$$

Q: why not restrict attention to x, x' that are neighbors?

Goal: add noise proportional to $\mathrm{RS}_q^{\sim,\mathcal{H}}$ when $x \in \mathcal{H}$ while maintaining privacy for all $x \in \mathcal{G}$.

Sensitivity of Graph Queries

query	range	$\mathbf{LS}_q^{\text{edge}}(G)$	$\mathbf{GS}_q^{\text{edge}}$	$\mathbf{LS}_q^{\text{node}}(G)$	$\mathbf{GS}_q^{\text{node}}$	$\mathbf{RS}_q^{\text{node},d}$
# edges	$[0, dn/2]$					
max-deg	$[0, d]$					
# triangles	$[0, \Theta(nd^2)]$					

$d = \mathrm{maxdeg}(G)$, typically $d \ll n$

Sensitivity of Graph Queries

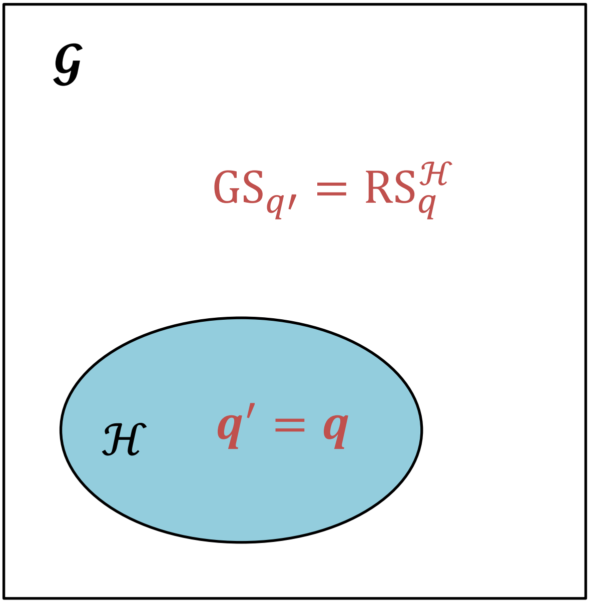
query	range	$\mathrm{LS}_q^{\mathrm{edge}}(G)$	$\mathrm{GS}_q^{\mathrm{edge}}$	$\mathrm{LS}_q^{\mathrm{node}}(G)$	$\mathrm{GS}_q^{\mathrm{node}}$	$\mathrm{RS}_q^{\mathrm{node},d}$
# edges	$[0, dn/2]$					
max-deg	$[0, d]$					
# triangles	$[0, \Theta(nd^2)]$					

$d = \mathrm{maxdeg}(G)$, typically $d \ll n$

Lipschitz extensions

Def: A function q' is a Lipschitz extension of q from $\mathcal{H}$ to $\mathcal{G}$ if

1. q' agrees with q on $\mathcal{H}$, and
2. $\mathrm{GS}_{q'} = \mathrm{RS}_q^{\mathcal{H}}$



- Can release q' via the Laplace Mechanism
- All real-valued functions have Lipschitz extensions [McShane 34]
- ∃ Lipschitz extensions for subgraph counts that can be computed efficiently [Kasiviswanathan-Nissim-Raskhodnikova-Smith `13]

[slide modified from one by Sofya Raskhodnikova]

Another Extension Theorem

Let:

- $\mathcal{G}$ = set of all datasets, with neighbor relation $\sim$.
- $d(x, x')$ = distance between datasets $x, x' \in \mathcal{G}$ wrt $\sim$
  = smallest k s.t. $\exists$ path $x = x_0 \sim x_1 \sim \cdots \sim x_k = x'$.
- $\mathcal{H} \subseteq \mathcal{G}$ be a "hypothesis" set of datasets in which we hope that our dataset lies.

Def: $M : \mathcal{H} \rightarrow \mathcal{Y}$ is ε-DP (restricted to $\mathcal{H}$) if $\forall\, x, x' \in \mathcal{H}, S \subseteq \mathcal{Y}$,

$$\Pr[M(x) \in S] \leq e^{\varepsilon \cdot d(x,x')} \cdot \Pr[M(x') \in S]$$

Thm [Borgs-Chayes-Smith-Zadik `18]: If $M : \mathcal{H} \rightarrow \mathcal{Y}$ is ε-DP (restricted to $\mathcal{H}$) then there is M': $\mathcal{G} \rightarrow \mathcal{Y}$ such that:

- M' is 2ε-DP, and
- For all $x \in \mathcal{H}$, $M(x) \equiv M(x')$.

Takeaways on Graph Privacy

- So far, we have focused on DP for “flat tables” where one individual’s data = one row.

- DP can be defined for graph data, and is compatible with many interesting statistical analyses.

- The theory of DP for graphs has been developed more recently and has not yet made its way into practice (as far as we know), but it may be ripe for doing so.