香港中文大學(深圳)理工学院
School of Science and Engineering

CSC4008: Techniques for Data Mining

Project Introduction

Jan. 28, 2021

Chenye Wu

wuchenye@cuhk.edu.cn

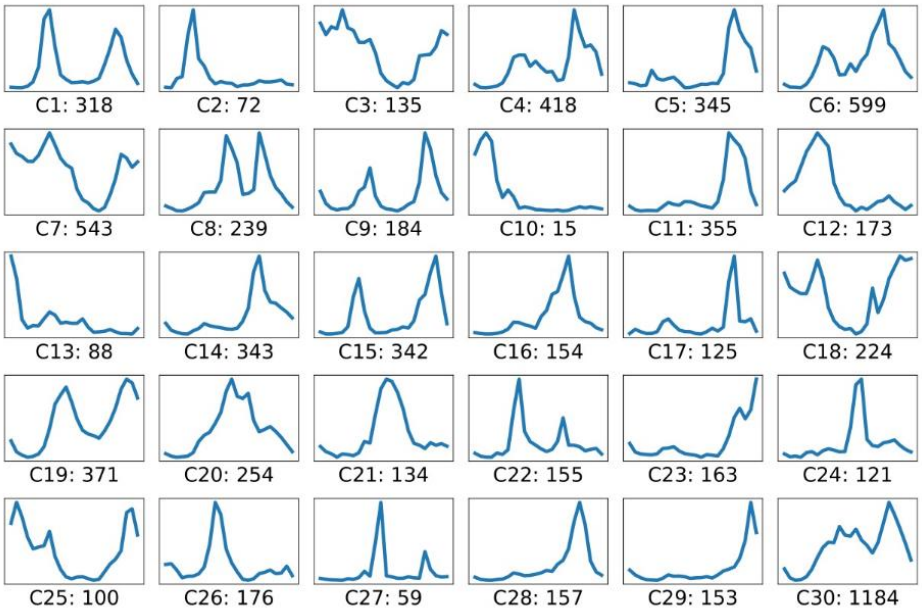
Example 3:

Vulnerability of Data-driven Pricing

Jingshi Cui, Haoxiang Wang,Chenye Wu, Yang Yu, “Robust Data-driven Profile-based Pricing Schemes", In Proceedings of IEEE ISGT NA 2021, Washington D.C. USA

Clustering Results

- Using Pecan Street Dataset
  - K-means clustering using l_1-norm!



Why choose l_1-norm?

- Cost function
  - $C_T(\boldsymbol{L}) = \sum_{t=1}^{T} C(L_t) = \sum_{t=1}^{T} \frac{1}{2} a \cdot {L_t}^2 + b \cdot L_t + c, \forall\, 1 \le t \le T, \quad s.t.\ L_t = \sum_{i=1}^{N} l_i^t$
- Real time price (real time marginal cost)
  - $p(t) = \frac{\partial C(L_t)}{\partial L_t} = a \cdot L_t + b, \forall\, 1 \le t \le T$
- Marginal System Cost Impact (MCI)
  - $MCI_i = \lim_{\Delta \to 0} \frac{C_T\left(\boldsymbol{L} + \Delta \frac{\boldsymbol{l_i}}{\|\boldsymbol{l_i}\|_1}\right) - C(\boldsymbol{l})}{\Delta} = \sum_{t=1}^{T} \frac{l_i^t}{\sum_{m=1}^{T} l_i^m} p(t)$

If we implement MCI as a uniform price, why it is good?

What does MCI tell us?

- If we choose to implement a uniform price, then we should choose MCI.
- And MCI is uniquely determined by user's load profiles.
- One straightforward implementation can be to conduct the k-means clustering based on load profiles.
- Users in the same cluster share the same price.

Clustering for Pricing

- **Clustering Challenges**
  - Problem: users in the same cluster do not share exactly the same load profile.
  - Loophole: certain users may bypass to other clusters for a better retail price.

Potential Solutions

- **Two Questions**
  - What are the possible strategic behaviors?
  - How many users can conduct price manipulation?
- **Our Solution**
  - Defining disguising ⇒ to identify strategic behavior.
  - Characterizing sensitive zone ⇒ to observe the impacts of different parameters.
  - Based on sensitive zone characterization and cost benefit analysis ⇒ vulnerability analysis.

Disguising: Strategic Behaviors

- Assumption
  - All users know the global information (central profile of each cluster and its corresponding price).
- Minimal Effort for Disguising

$$min_{n \neq u(i)}\ inf \lambda_{i,n}$$

user i's effort to move to cluster n

$$s.t.\ \left|\left|\left(1-\lambda_{i,n}\right)\boldsymbol{d}_i + \lambda_{i,n}\boldsymbol{c}_n - \boldsymbol{c}_{u(i)}\right|\right|_1 \geq \left|\left|\left(1-\lambda_{i,n}\right)\left(\boldsymbol{d}_i - \boldsymbol{c}_n\right)\right|\right|_1$$

$$p_n < p_{u(i)}$$

successfully switch to cluster n & lower price

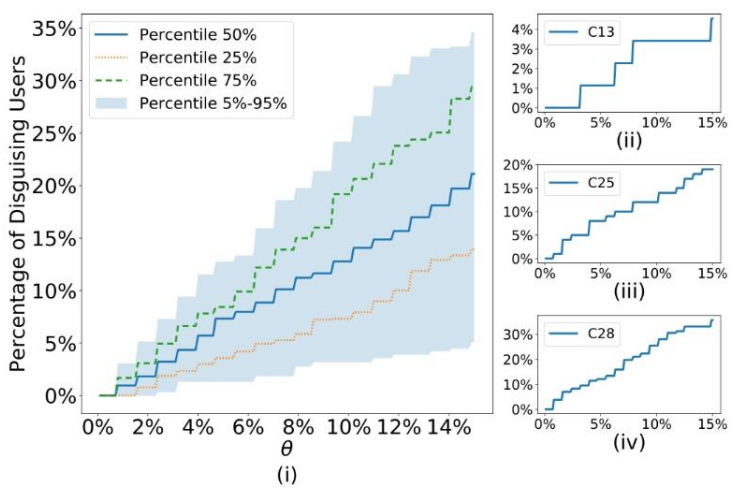
- Index to Differentiate Disguising
  - $CR_i = min_{n \in \{1,\ldots,k\}, n \neq u(i)}\ inf \lambda_{i,n}$

$u(i)$ is the cluster that user i belongs to.

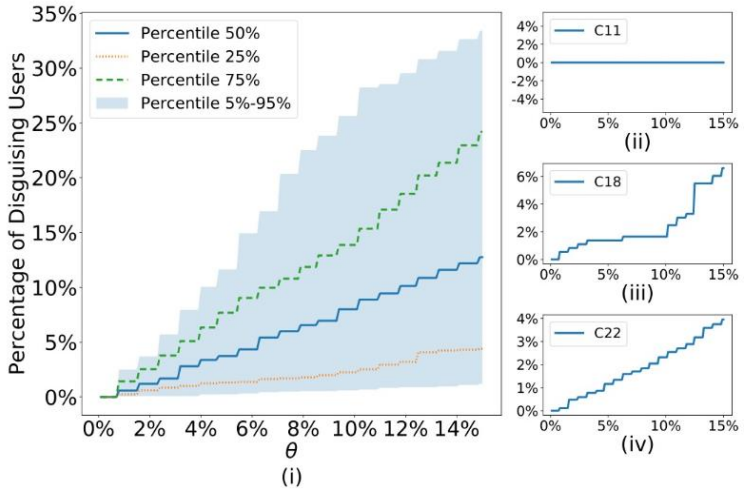
- Parametric Definition of Disguising
  - $CR_i \leq \theta$

Empirical Evidence: Strategic Behaviors

- Index to Quantify the Existence of Strategic Behaviors
  - $N_n(\theta) = \sum_{i,u(i)=n} I(CR_i \leq \theta)$
    - To define the number of users, having the ability to disguise, in each cluster n.



(a) Residential Users



(b) Commercial Building Users

Why Such Strategic Behaviors Exist?

- From End-to-End machine learning perspective…
  - It's because we select the wrong clustering criteria.
- We conduct k-means clustering for pricing. Hence, we should directly conduct the k-means clustering based on MCI.

- We define smoothness to guarantee robustness.
- ***(Definition)*** The $k-$means clustering is $\delta-$smooth if for any user i and its associated disguising set $\mathcal{K}_i$, the following condition holds:

$$|p_{u(i)} - p_n| \leq \delta, \qquad \forall\, n \in \mathcal{K}_i$$

Local properties for smoothness!

- ***k*−means Clustering with Smoothness Guarantee**
  - ***(Theorem)*** Suppose a *k*−means clustering algorithm can guarantee that

$$|MCI_i - p_{u(i)}| \leq \rho, \ \forall \ i \in u(i),$$

then, the clustering is $\rho(1 + \frac{1}{1-\theta})$-smooth.

A Greedy k-means Clustering is Optimal

Algorithm 2 Greedy $k-$means Clustering (GkC)

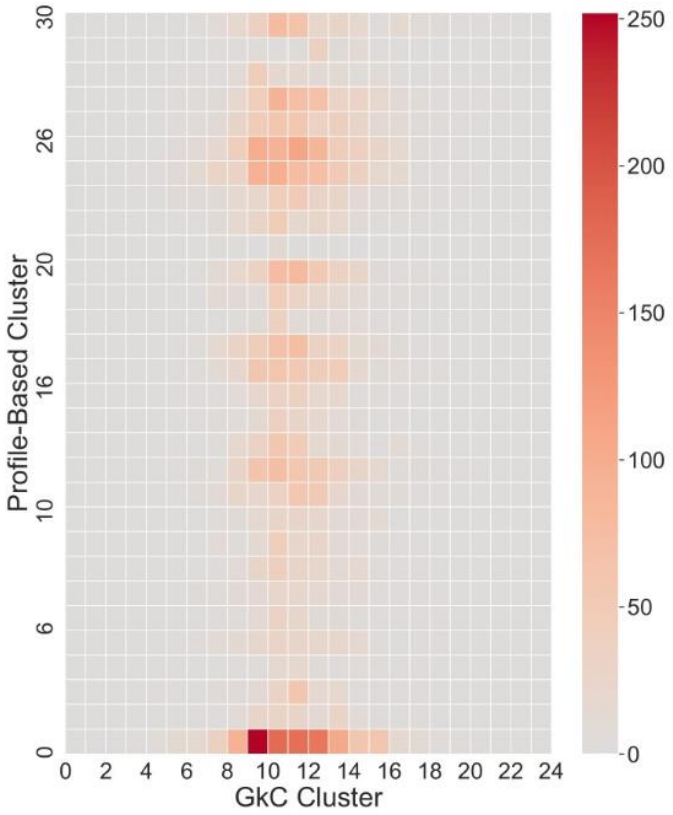
Input: Ordered tuple $(i, MCI_i), i = 1, ..., n$
Output: Clusters $C_1, ..., C_\kappa$

$i \leftarrow 1$
$k \leftarrow 1$
while $i \neq n$ **do**
  $s = MCI_i$
  $j = \arg\max_j \{MCI_j \leq MCI_i + 2\rho\}$
  $e = MCI_j$
  $C_k = \{i, ..., j\}$;
  $k \leftarrow k + 1$
  $i \leftarrow j + 1$
end while
$\kappa = k$;
return $C_1, ..., C_\kappa$

- **Optimality of G*k*C**
  - ***(Theorem)*** The G*k*C selects the minimal number of clusters to ensure the clustering criteria:

$$|MCI_i - p_{u(i)}| \leq \rho, \qquad \forall\, i$$

Existence of loopholes in the market…



Example 4:

Stability of k-means Clustering

Privacy is important

- We don’t want others to decipher our lifestyle…
- It can be dangerous…
- Texas Pressure Cooker Event

What we can do…

- We can inject noise into the metered data to achieve certain level of privacy.

- The most classical definition for privacy is differential privacy.

Definition 1. If for all neighbor datasets D_1 and D_2, and for all measurable subsets $Y \subset \mathbb{R}$, the mapping $\mathscr{B}$ satisfies,

$$\frac{Pr(\mathscr{B}(D_1) \in Y)}{Pr(\mathscr{B}(D_2) \in Y)} \leq e^{\epsilon}, \tag{3}$$

we say the mechanism $\mathscr{B}$ achieves ϵ-DP [28].

How to achieve DP?

Theorem 1. We say a mechanism $\mathscr{B}$ achieves ϵ-DP, if $\mathscr{B}(D)$ satisfies

$$\mathscr{B}(D) = q(D) + n, \tag{4}$$

and n is Laplace noise with probability density function (pdf) $p(s)$:

$$p(s) = \frac{1}{2\lambda} e^{-\frac{|s|}{\lambda}}, \tag{5}$$

in which $\lambda = \frac{\Delta f(D)}{\epsilon}$ combines the privacy parameter ϵ that describes the level of the privacy and the sensitivity $\Delta f(D)$ satisfying $\Delta f(D) \geq \max \|q(D_1) - q(D_2)\|$ for all neighbor datasets D_1 and D_2.

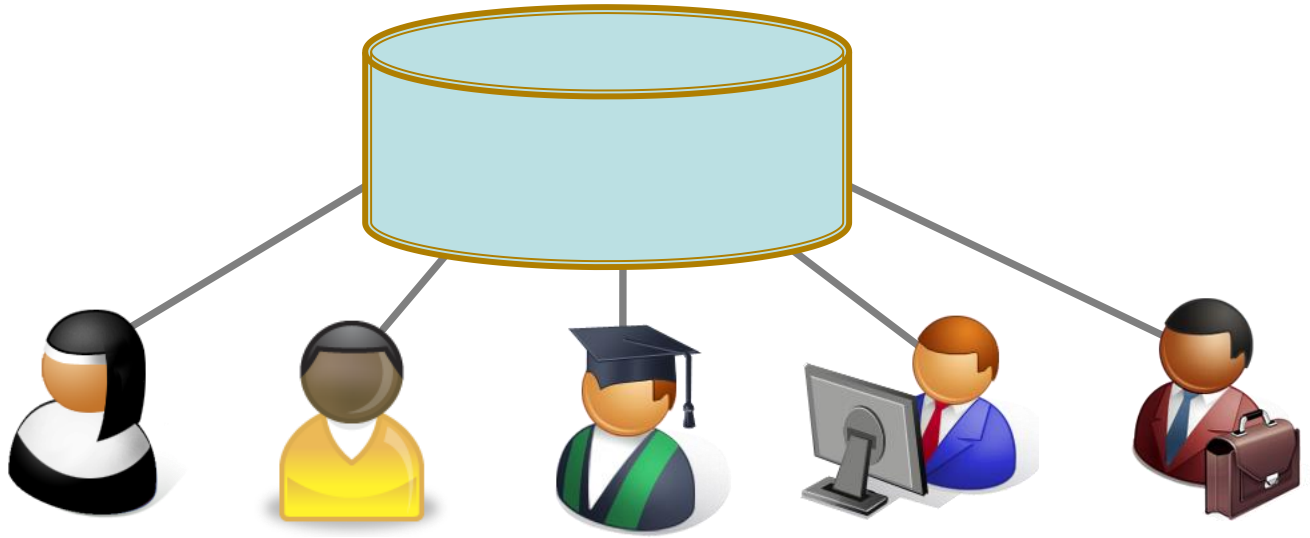
Intuition for Differential Privacy

Some slides adapted from Adam Smith & Elaine Shi's lecture and other talk slides

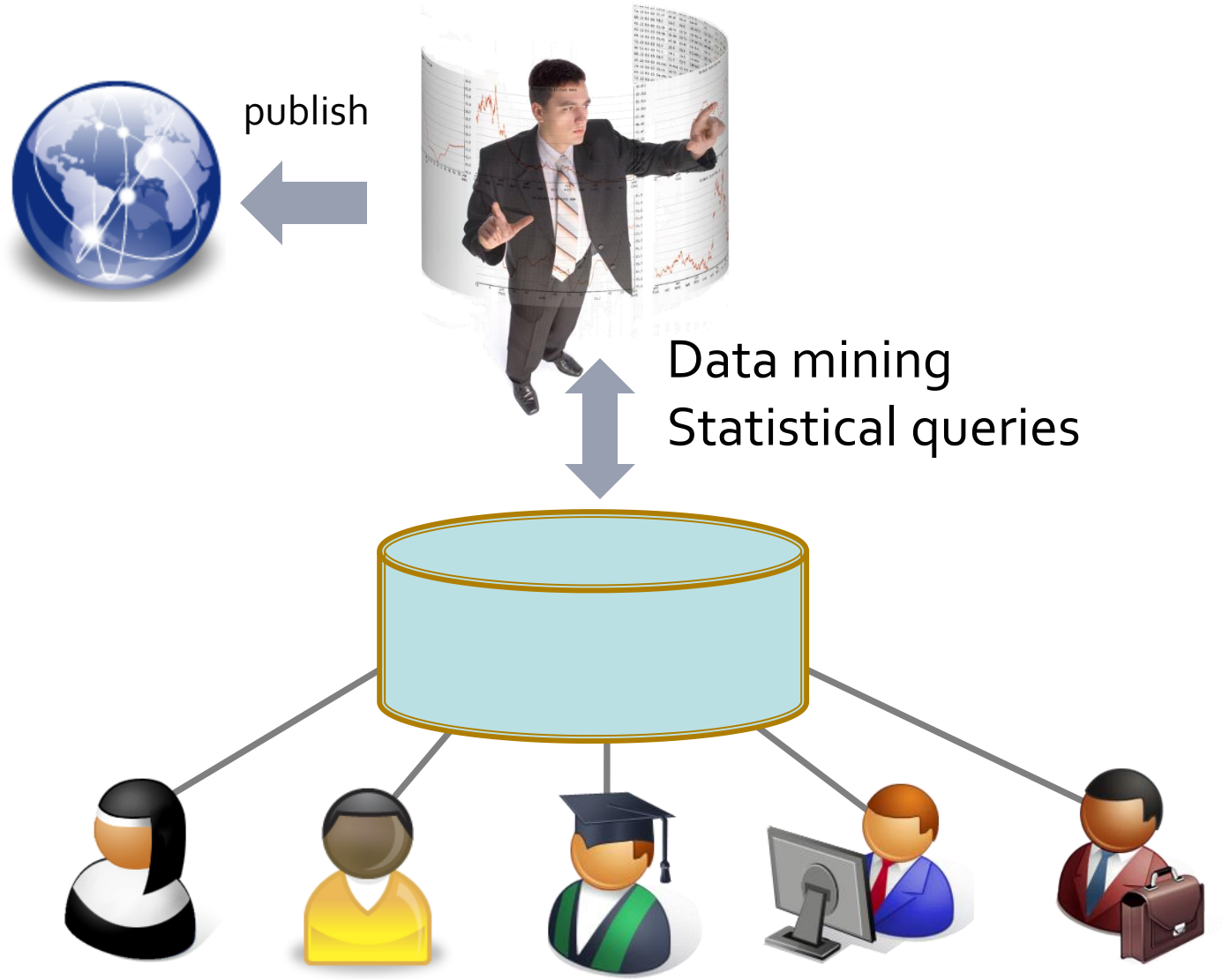
General Setting

Medical data
Query logs
Social network data
…

Data mining
Statistical queries



General Setting



How can you allow meaningful usage of such datasets while preserving individual privacy?

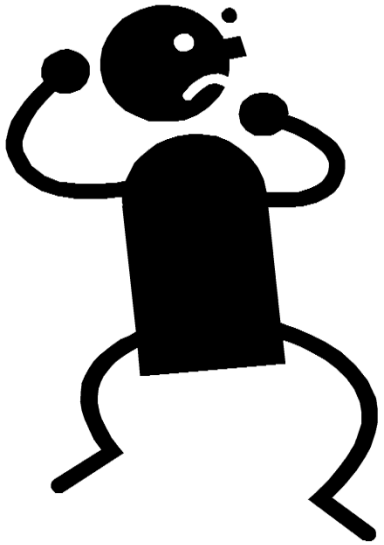
Blatant Non-Privacy

Blatant Non-Privacy

- Leak individual records
- Can link with public databases to re-identify individuals
- Allow adversary to reconstruct database with significant probablity

Attempt 1: Crypto-ish Definitions

I am releasing some useful statistic f(D), and nothing more will be revealed.

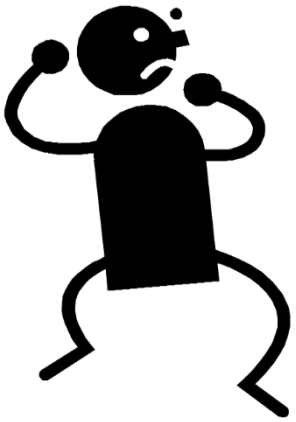


What kind of statistics are safe to publish?

How do you define privacy?

Attempt 2:

I am releasing researching findings showing that people who smoke are very likely to get cancer.



You cannot do that, since it will break my privacy. My insurance company happens to know that I am a smoker…

Attempt 2: Absolute Disclosure Prevention

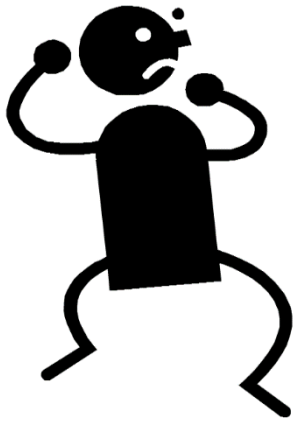
"If the release of statistics *S makes it possible to* determine the value [of private information] more accurately than is possible without access to *S*, a disclosure has taken place."
[Dalenius]

An Impossibility Result

[informal] It is not possible to design any non-trivial mechanism that satisfies such strong notion of privacy.[Dalenius]

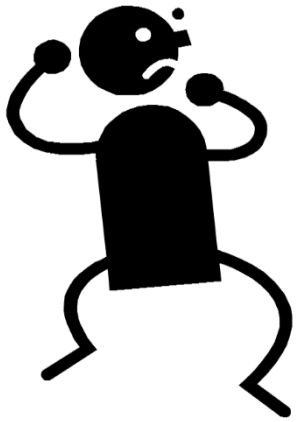
Attempt 3: "Blending into Crowd" or k-Anonymity

K people purchased A and B, and all of them also purchased C.



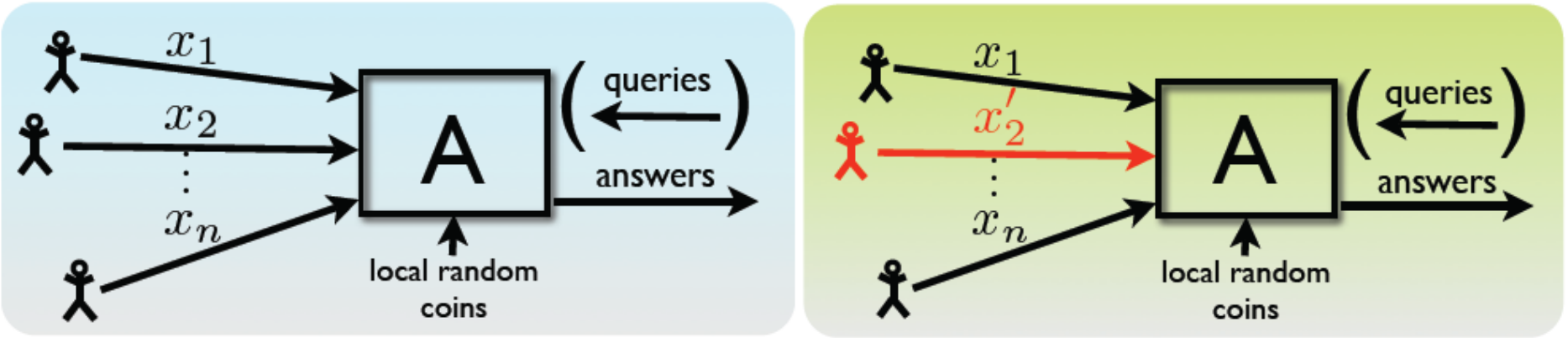
Attempt 3: "Blending into Crowd" or k-Anonymity

K people purchased A and B, and all of them also purchased C.



I know that Elaine bought A and B…

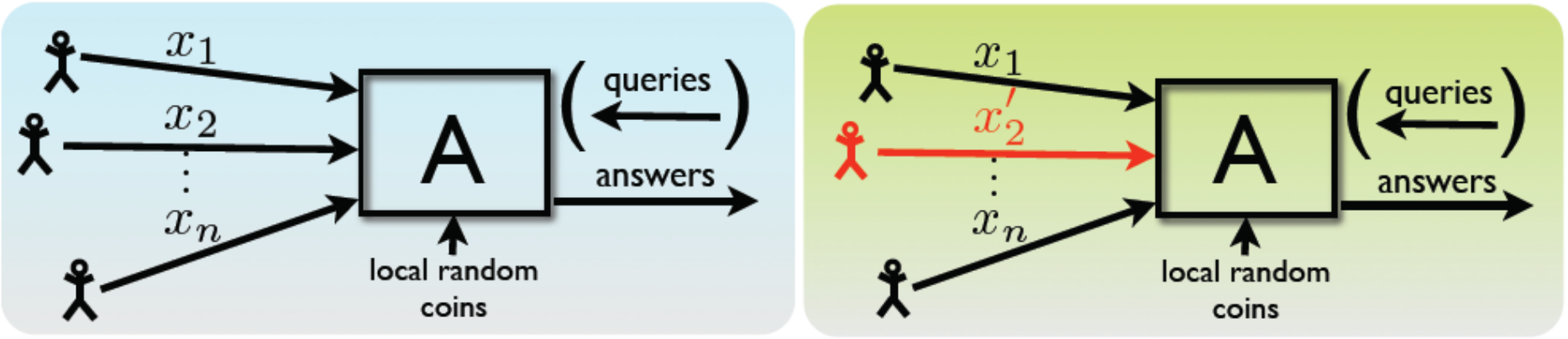
Attempt 4: Differential Privacy



x' is a neighbor of x
if they differ in one row

From the released statistics, it is hard to tell which case it is.

Attempt 4: Differential Privacy



x' is a neighbor of x
if they differ in one row

For all neighboring databases x and x'
For all subsets of transcripts:

$$\Pr[A(x) \in S] \leq e^{\epsilon} \Pr[A(x') \in S]$$

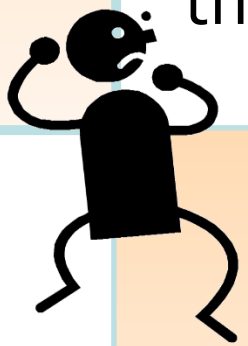
Attempt 4: Differential Privacy

I am releasing researching findings showing that people who smoke are very likely to get cancer.

1

Please don't blame me if your insurance company knows that you are a smoker, since I am doing the society a favor.

2



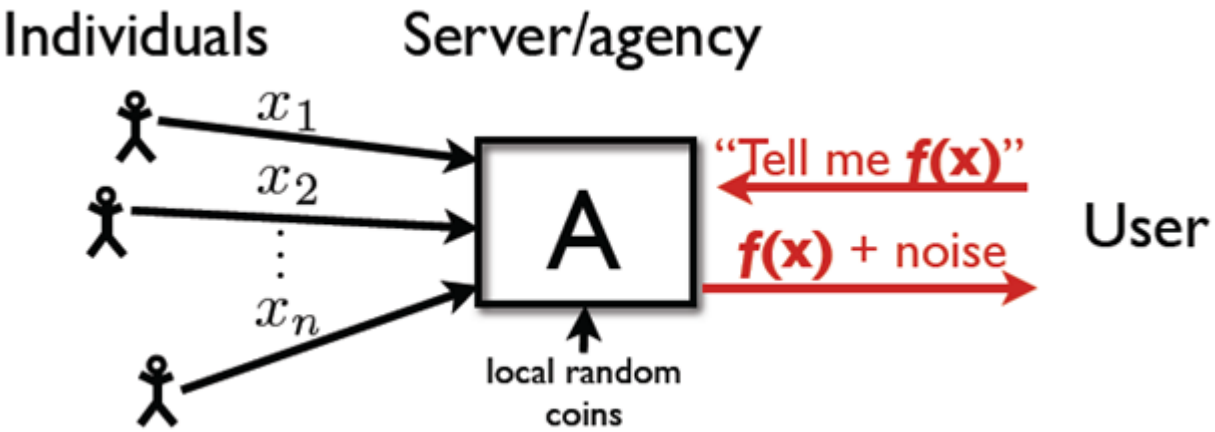
Oh, btw, please feel safe to participate in my survey, since you have nothing more to lose.

3

Since my mechanism is DP, whether or not you participate, your privacy loss would be roughly the same!

4

Method1: Output Perturbation



- Global Sensitivity:

$$GS_f = max\|f(x) - f(x')\|_1$$

x,x′ neighbors

Example: $GS_{avg} = \frac{1}{n}$

One Method: Output Perturbation

- Theorem:

$$A(x) = f(x) + Lap(\frac{GS_f}{\epsilon}) \text{ is } \epsilon\text{-DP}$$

- Intuition: add more noise when function is sensitive

How to evaluate DP's impact?

- Stability of Clustering!!!

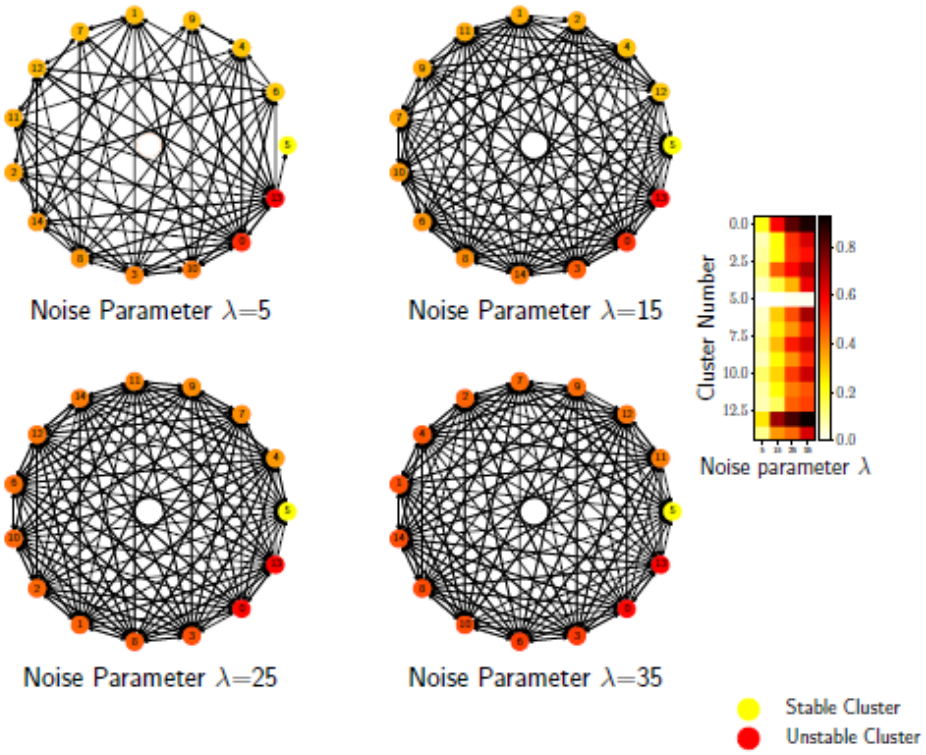


Fig. 3. The clusters interchanges influenced by different levels of noise injection

Math Genius Could Prove Bounds…

Theorem 2. If Assumption 1 holds, the probability that the consumer would stay in his original cluster after Laplace noise injection (with parameter λ) is lower bounded by

$$C_0(r)\lambda^{-T+1}\exp\left(\frac{-C(r)}{\lambda}\right), \tag{8}$$

where C_0 and C are constants related to the radius r.

Roadmap of the Proof

We fix $\lambda = \frac{\delta}{2\epsilon}$ according to Theorem 1. Denote the original load profile by $\boldsymbol{d}$, and then we could use a normalized vector $\boldsymbol{s} \in \mathbb{R}^T$ to define the pdf $q(\cdot)$ for the normalized demand after noise injection:

$$q(\boldsymbol{s}) = \int_0^{+\infty} \frac{k^{T-1}}{(2\lambda)^T} \exp\left(-\frac{\|k\mathbf{s} - \mathbf{d}\|_1}{\lambda}\right) dk \tag{9}$$

Then we denote the cluster stability by PL. Specifically, if the profile vector $\boldsymbol{s}$ such that $\|\boldsymbol{s} - \tilde{\boldsymbol{d}}_0\|_2 \leq r$, the profile belongs to cluster 0. This is due to the following triangle inequality:

$$\begin{aligned} \|\boldsymbol{s} - \tilde{\boldsymbol{d}}_i\|_2 &\geq \left| \|\boldsymbol{s} - \tilde{\boldsymbol{d}}_0\|_2 - \|\tilde{\boldsymbol{d}}_i - \tilde{\boldsymbol{d}}_0\|_2 \right| \\ &\geq r \geq \|\boldsymbol{s} - \tilde{\boldsymbol{d}}_0\|_2 \end{aligned} \tag{10}$$

Thus, define $\Theta = \{\boldsymbol{s} | \|\boldsymbol{s} - \tilde{\boldsymbol{d}}_0\|_2 \leq r, \|\boldsymbol{s}\|_2 = 1\}$ as the cluster region for $\tilde{\boldsymbol{d}}_0$. We could derive the lower bound for PL by $PL \geq \int_\Theta q(\boldsymbol{s}) d\boldsymbol{s}$. We need to define a constant τ for all $\boldsymbol{s}$ in Θ, such that $\tau \boldsymbol{s} \geq \boldsymbol{d}$ for all dimensions.

Lower Bound for q(s)

$$
\begin{aligned}
q(\boldsymbol{s}) &= \frac{1}{(2\sqrt{T})^T} \exp\left(\frac{\|\mathbf{d}\|_1}{\lambda}\right) \times \\
&\int_{\frac{\tau\sqrt{T}}{\lambda}}^{+\infty} \frac{(k\sqrt{T})^{T-1}}{\lambda^{T-1}} \exp\left(-\frac{k\sqrt{T}}{\lambda}\right) d\left(\frac{k\sqrt{T}}{\lambda}\right) \\
&+ \int_0^{\tau} \frac{k^{T-1}}{(2\lambda)^T} \exp\left(-\frac{\|k\mathbf{s}-\mathbf{d}\|_1}{\lambda}\right) dk \qquad (11) \\
\geq &\frac{1}{(2\sqrt{T})^T} \exp\left(\frac{\|\mathbf{d}\|_1}{\lambda}\right) \left(\frac{\tau\sqrt{T}}{\lambda}\right)^{T-1} \exp\left(-\frac{\tau\sqrt{T}}{\lambda}\right) \\
&+ \int_0^{\tau} \frac{k^{T-1}}{(2\lambda)^T} \exp\left(-\frac{\|k\mathbf{s}\|_1 + \|\mathbf{d}\|_1}{\lambda}\right) dk
\end{aligned}
$$

Triangle inequality

Characteristics of Incomplete Gamma Function

$$
\begin{aligned}
q(\boldsymbol{s}) &\geq \frac{\tau^{T-1}}{2^T \lambda^{T-1}\sqrt{T}} \exp\left(\frac{\|\mathbf{d}\|_1 - \tau\sqrt{T}}{\lambda}\right) \\
&+ \frac{1}{(2\sqrt{T})^T} \exp\left(\frac{-\|\mathbf{d}\|_1}{\lambda}\right) \times \\
&\int_0^{\frac{\tau\sqrt{T}}{\lambda}} \frac{(k\sqrt{T})^{T-1}}{\lambda^{T-1}} \exp\left(-\frac{k\sqrt{T}}{\lambda}\right) d\left(\frac{k\sqrt{T}}{\lambda}\right) \\
&\geq \frac{\tau^{T-1}}{2^T \lambda^{T-1}\sqrt{T}} \exp\left(\frac{\|\mathbf{d}\|_1 - \tau\sqrt{T}}{\lambda}\right) \\
&+ \frac{1}{(2\sqrt{T})^T} \exp\left(\frac{-\|\mathbf{d}\|_1}{\lambda}\right) \frac{(\frac{\tau\sqrt{T}}{\lambda})^T}{T} \exp\left(\frac{-T\sqrt{T}\tau}{\lambda(T+1)}\right) \\
&= \frac{\tau^{T-1}}{2^T \lambda^{T-1}\sqrt{T}} \exp\left(\frac{\|\mathbf{d}\|_1 - \tau\sqrt{T}}{\lambda}\right) \\
&+ \frac{\tau^T}{(2\lambda)^T T} \exp\left(-\frac{\|\mathbf{d}\|_1}{\lambda} - \frac{T\sqrt{T}\tau}{\lambda(T+1)}\right)
\end{aligned}
$$

We know the volume of unit ball

$$
\begin{aligned}
PL &\geq \int_{\Theta} q(\boldsymbol{s}) d\boldsymbol{s} \\
&\geq \frac{\tau^{T-1}\pi^{\frac{T-1}{2}} r^{T-1}}{2^T \lambda^{T-1}\sqrt{T}\Gamma(\frac{T+1}{2})} \exp\left(\frac{\|\mathbf{d}\|_1 - \tau\sqrt{T}}{\lambda}\right) \\
&+ \frac{\tau^{T}\pi^{\frac{T-1}{2}} r^{T-1}}{(2\lambda)^T T\Gamma(\frac{T+1}{2})} \exp\left(-\frac{\|\mathbf{d}\|_1}{\lambda} - \frac{T\sqrt{T}\tau}{\lambda(T+1)}\right) \\
&\geq C_0(r)\lambda^{-T+1} \exp\left(\frac{-C(r)}{\lambda}\right),
\end{aligned} \tag{13}
$$

where

$$
C_0 = \frac{\tau^{T-1}\pi^{\frac{T-1}{2}} r^{T-1}}{2^T\sqrt{T}\Gamma(\frac{T+1}{2})}, \text{ and } C = \tau\sqrt{T} - \|\mathbf{d}\|_1. \tag{14}
$$

Utilize the Result

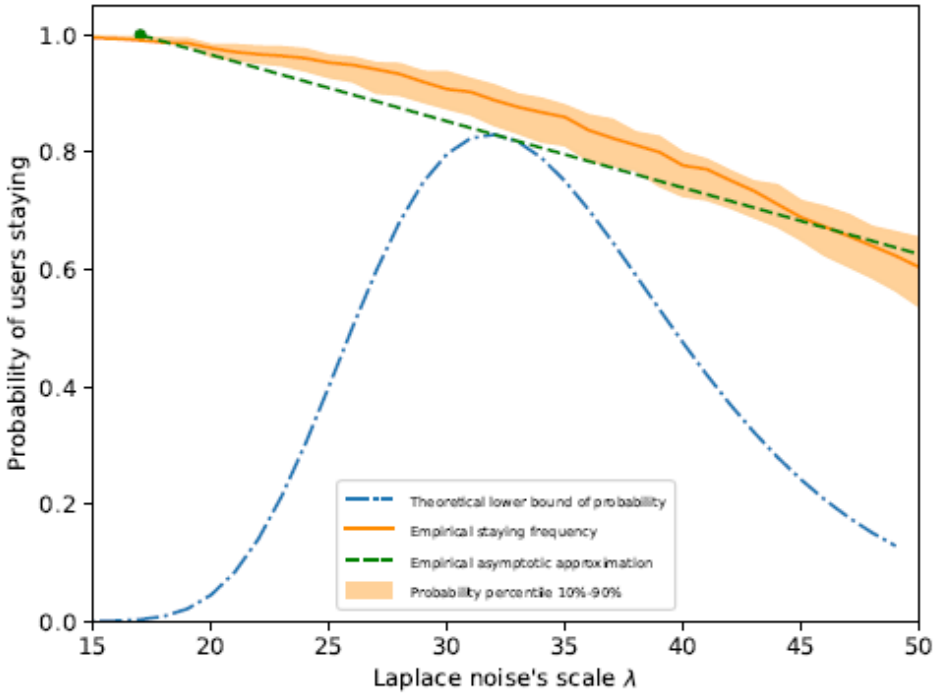


Fig. 4. The relationship between λ and staying probability