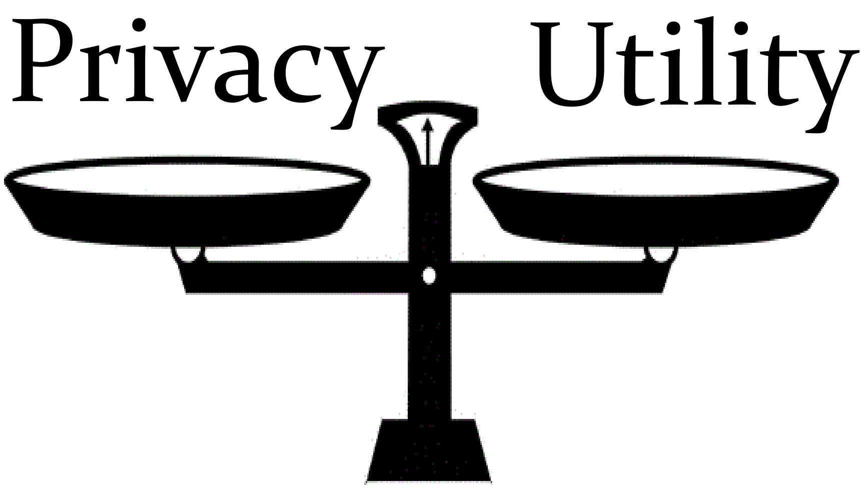
Mobile Data Collection and Analysis with Local Differential Privacy - Part 1

Ninghui Li (Purdue University)

Outline

- Motivation of Differential Privacy and Local Differential Privacy (LDP)
- Frequency Oracles in LDP

Tradeoff between Privacy and Utility



A **privacy notion** for privacy protection guarantee

Design a **mechanism** under such notion with high utility

AOL Data Release [NYTimes 2006]

- In August 2006, AOL Released search keywords of 650,000 users over a 3-month period.
  - User IDs are replaced by random numbers.
  - 3 days later, pulled the data from public access.

AOL searcher # 4417749
"landscapers in Lilburn, GA" *queries on last name "Arnold"* *"homes sold in shadow lake subdivision Gwinnett County, GA"* *"num fingers"* *"60 single men"* *"dog that urinates on everything"*

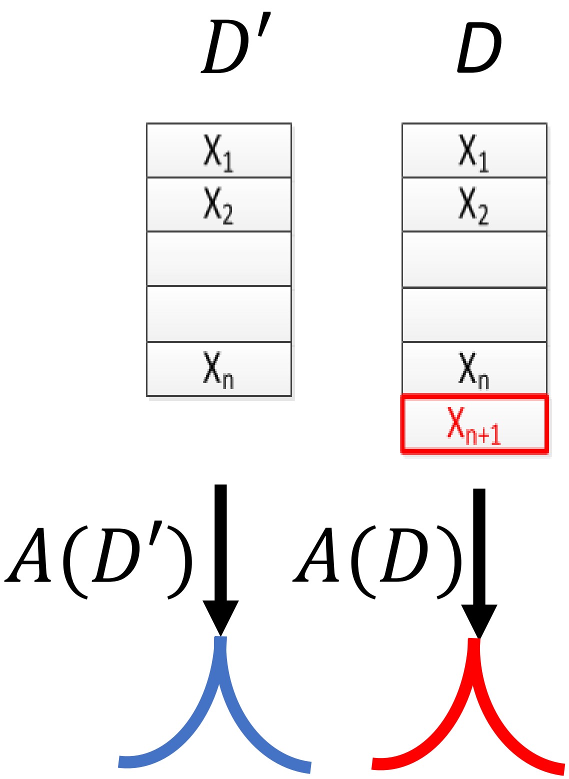
NYT

Thelman Arnold, a 62 year old widow who lives in Liburn GA, has three dogs, frequently searches her friends' medical ailments.

Re-identification occurs!

Differential Privacy [Dwork et al. 2006]

- Idea: Any output should be about as likely
  regardless of whether or not I am in the dataset



Def. Algo A satisfies ϵ-differential privacy if for any neighboring D and D’ and any possible output t,

$$e^{-\epsilon} \leq \frac{\Pr[A(D)=t]}{\Pr[A(D')=t]} \leq e^{\epsilon}$$

Parameter ϵ: strength of privacy protection, known as privacy budget.

Key Assumption Behind DP: The Personal Data Principle

- After removing one individual's data, that individual's privacy is protected perfectly.
    - Even if correlation can still reveal individual info, that is not considered to be privacy violation
- In other words, for each individual, the world after removing the individual's data is an ideal world of privacy for that individual. Goal is to simulate all these ideal worlds.

Differential Privacy in the Centralized Setting

Database

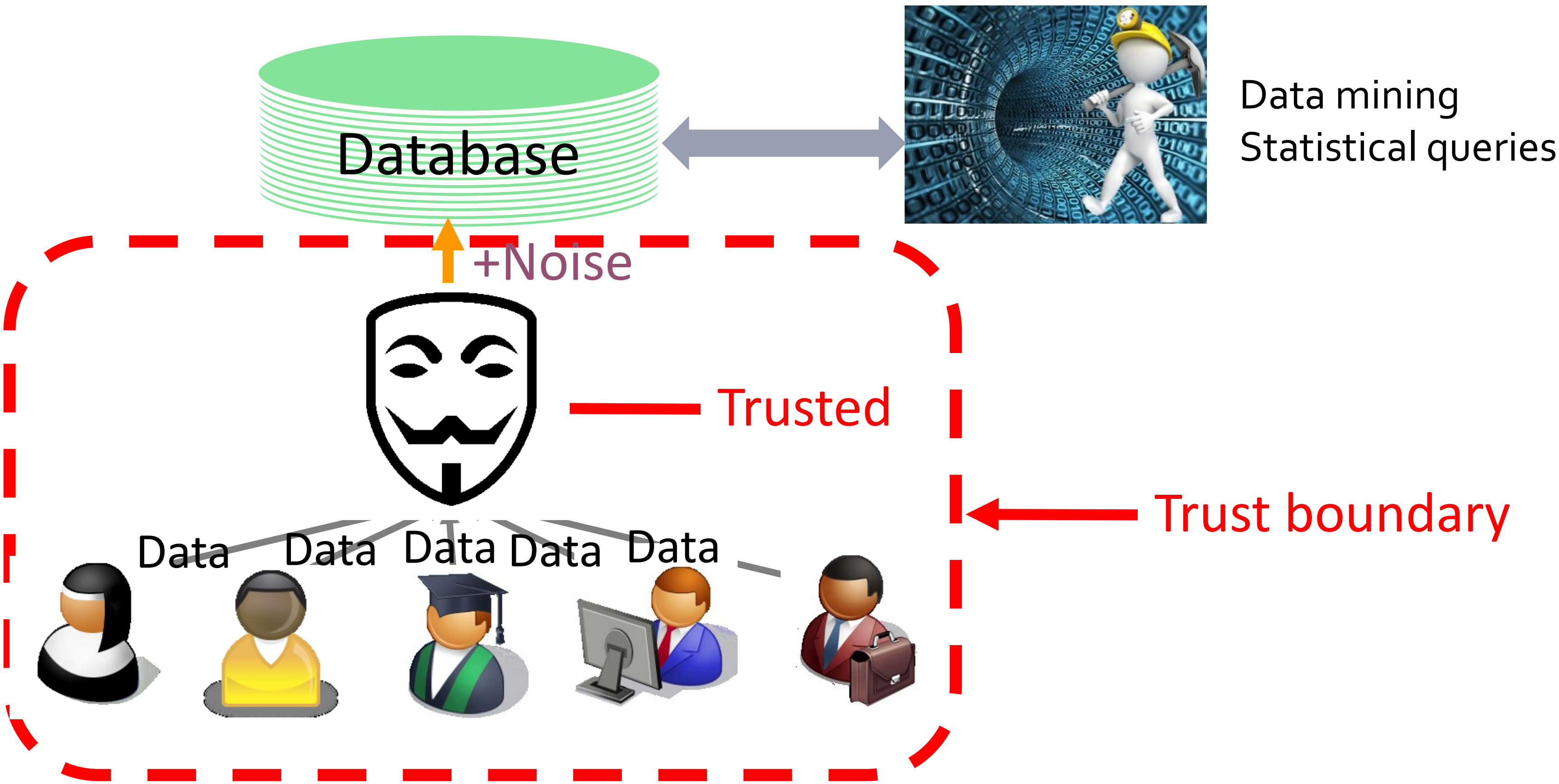
Data mining
Statistical queries

Classical/
centralized
setting

Differential Privacy Interpretation:
The decision to include/exclude an individual's record has limited (ε) influence on the outcome.
Smaller ε → Stronger Privacy

Data

Differential Privacy in the Centralized Setting



Local Differential Privacy

As Apple starts analyzing web browsing & health data, how comfortable are you with differential privacy?

RAPP

ng

s

Ben Lovejoy - Jul. 7th 2017 6:59 am PT @benlovejoy

er

Trust boundary

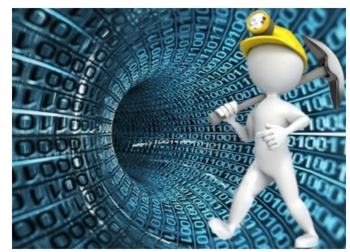
Outline

- Motivation of Differential Privacy and Local Differential Privacy (LDP)
- Frequency Oracles in LDP

The Frequency Oracle Protocols under LDP

- $y := P(v)$
takes input value v from domain D and outputs y.

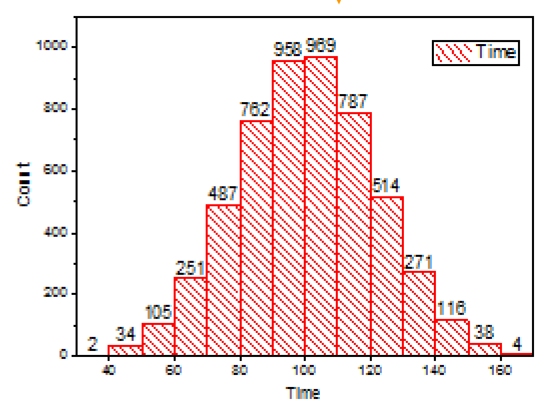
y



- $c := Est(\{y\})$
takes reports $\{y\}$ from all users and outputs estimations $c(v)$ for any value v in domain D

FO is ε -LDP iff for any v and v' from D, and any valid output y,

$$\frac{\Pr[P(v)=y]}{\Pr[P(v')=y]} \leq e^{\varepsilon}$$



Random Response (Warner'65)

- Survey technique for private questions
- Survey people:
  - "Do you a disease?"
- Each person:
  - Flip a secret coin
  - Answer truth if head (w/p 0.5)
  - Answer randomly if tail
  - E.g., a patient will answer "yes" w/p 75%, and "no" w/p 25%
- To get unbiased estimation of the distribution:
  - If n_v out of n people have the disease, we expect to see

    $E[\,I_v] = 0.75n_v + 0.25(n - n_v)$ "yes" answers
  - $c(n_v) = \frac{I_v - 0.25n}{0.75 - 0.5}$ is the unbiased estimation of number of patients

Provide deniability:
Seeing answer, not certain about the secret.

Concrete Example

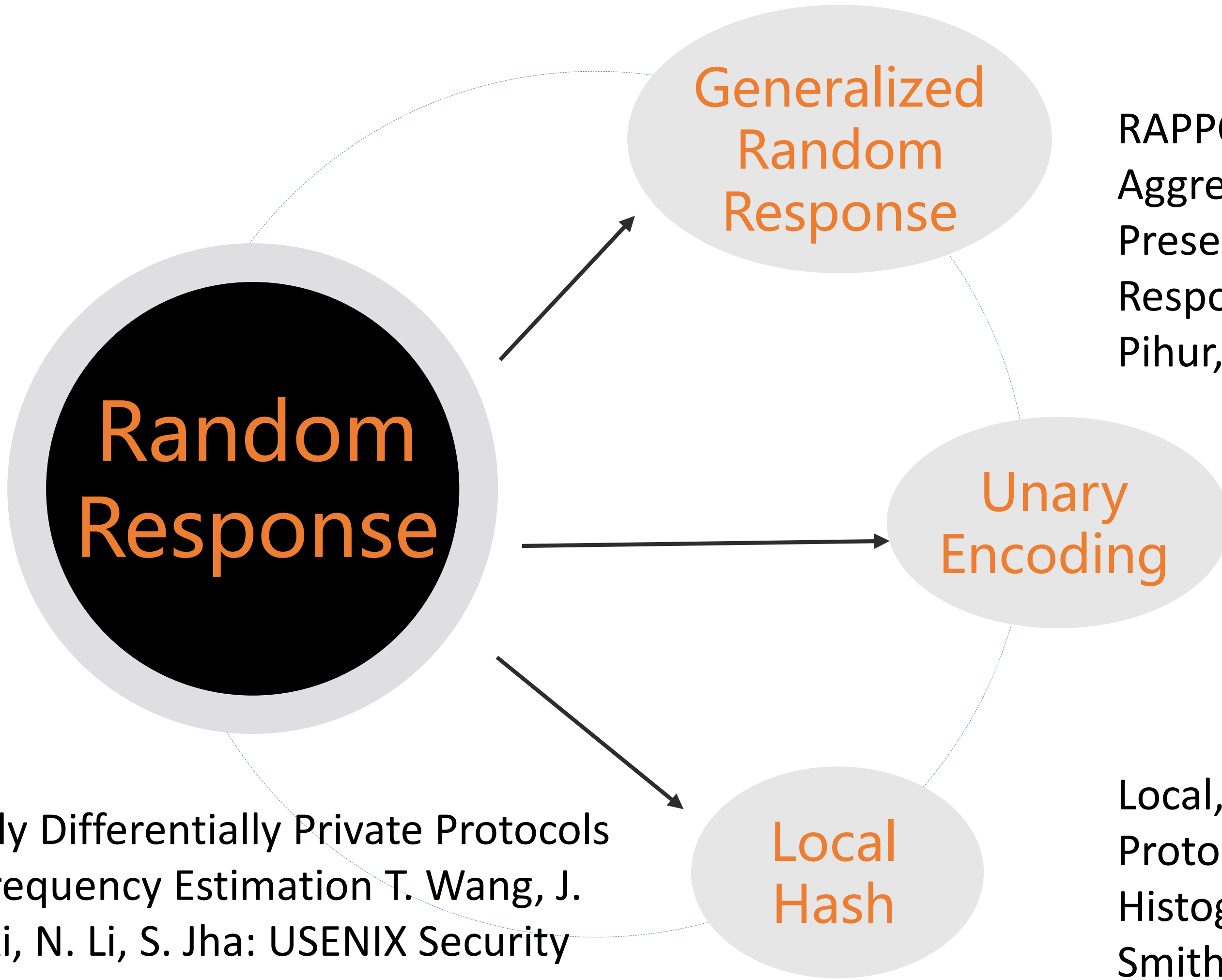
An individual will answer "yes" w/p 75%, and "no" w/p 25%

	truth	Expected yes	Expected no
yes	80	60	20
no	20	5	15

$$c(n_v) = \frac{I_v - 0.25n}{0.75 - 0.25}$$

observed	65	35
estimate	80	20

From Two to Any Categories



RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. Ú. Erlingsson, V. Pihur, A. Korolova, CCS 2014

Local, Private, Efficient Protocols for Succinct Histograms R. Bassily, A. Smith. STOC 2015.

Locally Differentially Private Protocols for Frequency Estimation T. Wang, J. Blocki, N. Li, S. Jha: USENIX Security 2017

Generalized Random Response

- User:
  - Given $v \in D = \{1,2,\ldots,d\}$
  - Toss a
  - If it is
  - Otherwise, report any other value with probability $q = \frac{1-p}{d-1}$

Intuitively, the higher p, the more accurate

However, when d is large, p becomes small (for the same ε)

ε	$p(d=2)$	$p(d=8)$	$p(d=128)$	$p(d=1024)$
0.1	0.52	0.13	0.016	0.001
1	0.73	0.27	0.027	0.002
2	0.88	0.51	0.057	0.007
4	0.98	0.88	0.307	0.05

  - $E[I_v] = n_v \cdot p + (n - n_v) \cdot q$
  - Unbias

To get rid of dependency on domain size, we move to the other protocols.

Unary Encoding (Basic RAPPOR)

- Encode the value v into a bit string $\boldsymbol{x} \coloneqq \vec{0}, \boldsymbol{x}[v] \coloneqq 1$
  - e.g., $D = \{1,2,3,4\}, v = 3,$ then $\boldsymbol{x} =$ [0,0,1,0]
- Perturb each bit, preserving it with probability p
  - $p_{1\to1} = p_{0\to0} = p = \frac{e^{\varepsilon/2}}{e^{\varepsilon/2}+1}$ $\quad p_{1\to0} = p_{0\to1} = q = \frac{1}{e^{\varepsilon/2}+1}$
  - $\Rightarrow \frac{\Pr[P(E(v))=\boldsymbol{x}]}{\Pr[P(E(v'))=\boldsymbol{x}]} \leq \frac{p_{1\to1}}{p_{0\to1}} \times \frac{p_{0\to0}}{p_{1\to0}} = e^{\varepsilon}$
    - Since $\boldsymbol{x}$ is unary encoding of v, $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ in two locations
- Intuition:
  - By unary encoding, each location can only be 0 or 1, effectively reducing d in each location to 2. (But privacy budget is halved.)
  - When d is large, UE is better than DE.
- To estimate frequency of each value, do it for each bit.

Binary Local Hash

- The original protocol uses a shared random matrix; this is an equivalent description
- Each user uses a random hash function from D to $\{0,1\}$
- The user then perturbs the bit with probabilities
  - $p = \frac{e^{\varepsilon}}{e^{\varepsilon}+1}, q = \frac{1}{e^{\varepsilon}+1}$

$$\Rightarrow \frac{\Pr[P(E(\boldsymbol{v})) = b]}{\Pr[P(E(\boldsymbol{v}')) = b]} = \frac{p}{q} = e^{\varepsilon}$$

- The user then reports the bit and the hash function
- The aggregator increments the reported group
- $E[I_v] = n_v \cdot p + (n - n_v) \cdot (\frac{1}{2}q + \frac{1}{2}p)$
- Unbiased Estimation: $c(v) = \frac{I_v - n \cdot \frac{1}{2}}{p - \frac{1}{2}}$

Optimization

- We measure utility of a mechanism by its variance
  - E.g., in Random Response,
    - $Var[c(v)] = Var\left[\frac{I_v - n\cdot q}{p-q}\right] = \frac{Var[I_v]}{(p-q)^2} \approx \frac{n\cdot q\cdot(1-q)}{(p-q)^2}$
- We propose a framework called ‘pure’ and cast existin
  - Eac
    - orts each
    - E.g., in BLH, Support(y) = $\{v \mid H(v) = y\}$
  - A pure protocol is specified by p' and q'
    - Each input is perturbed into a value “supporting it” with p', and into a value not supporting it with q'

$min_{q'} Var[c(v)]$

or $min_{q'} \frac{n\cdot q'\cdot(1-q')}{(p'-q)'^2}$

where p', q' satisfy ε-LDP

Frequency Estimation Protocols

- Randomised response: a survey technique for eliminating evasive answer bias
  - S.L. Warner, Journal of Ame. Stat. Ass. 1965
  - Direct Encoding (Generalized Random Response)
- RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response.
  - Ú. Erlingsson, V. Pihur, A. Korolova, CCS 2014
  - Unary Encoding, Encode into a bit-vector
- Local, Private, Efficient Protocols for Succinct Histograms
  - R. Bassily, A. Smith. STOC 2015.
  - Binary Local Hash: Encode by hashing and then perturb
- Locally Differentially Private Protocols for Frequency Estimation
  - T. Wang, J. Blocki, N. Li, S. Jha: USENIX Security 2017

Optimized Local Hash (OLH)

- In original BLH, secret is **compressed** into a bit, **perturbed** and transmitted.
- Both steps cause information loss:
  - Compressing: loses much
  - Perturbation: information loss depends on ϵ
- **Key Insight**: We want to make a balance between the two steps:
  - By compressing into more groups, the first step carries more information
- Variance is optimized when $g = e^{\varepsilon} + 1$
- See our paper for details.

Other Topics

- Dearling with numerical data, estimating mean:
  - Goal: Find the mean of continuous values
  - Assumption: Each user has a single value x within the range of $[-1, +1]$
  - Intuition: Report +1 with higher probability if x closer to +1
  - [https://arxiv.org/abs/1606.05053,https://arxiv.org/pdf/1712.01524]
- Frequent itemset mining:
  - Zhan Qin, et al.: Heavy Hitter Estimation over Set-Valued Data with Local Differential Privacy. ACM CCS 2016
  - Tianhao Wang, Ninghui Li, Somesh Jha: **Locally Differentially Private Frequent Itemset Mining.** IEEE Symposium on Security and Privacy 2018

Other interesting problems

- Stochastic gradient descent
  - Goal: Find the optimal machine learning model
  - Assumption: Each user has a vector $\boldsymbol{x}$
  - Intuition: Bolt-on sgd with noisy update
  - [https://arxiv.org/abs/1606.05053]
- Bound the privacy leakage
  - Goal: Make multiple, periodic collection possible
  - Assumption: Each user has a value $x(t)$ that change with time
  - Intuition: Decide whether to participate based on the current result
  - [https://arxiv.org/abs/1802.07128]
- Many more

Mobile Data Collection and Analysis with Local Differential Privacy - Part 2

Qingqing Ye

Renmin University of China

Hong Kong Polytechnic University

Outline

- Current Research Problem
    - Marginal Release
    - Graph Data Mining
    - Key-Value Data Collection

- Open Problems and New Directions
    - Iterative Interaction
    - Privacy-Preserving Machine Learning
    - Theoretical underpinning

Outline

- Current Research Problem
  - **Marginal Release**
  - Graph Data Mining
  - Key-Value Data Collection

- Open Problems and New Directions
  - Iterative Interaction
  - Privacy-Preserving Machine Learning
  - Theoretical underpinning

Marginal Release

- Full contingency table: distribution of all attribute combinations

Dataset:

User	Gender	Smoke
Alice	female	smoker
Bob	male	non-smoker
Tom	male	smoker
…		
Lily	female	non-smoker

2-way marginal

v	$F(v)$
< female, non-smoker >	0.35
< female, smoker >	0.15
< male, non-smoker >	0.1
< male, smoker >	0.4

- Marginal table: distribution of part of attribute combinations

v	$F(v)$
< female, * >	0.5
< male, * >	0.5

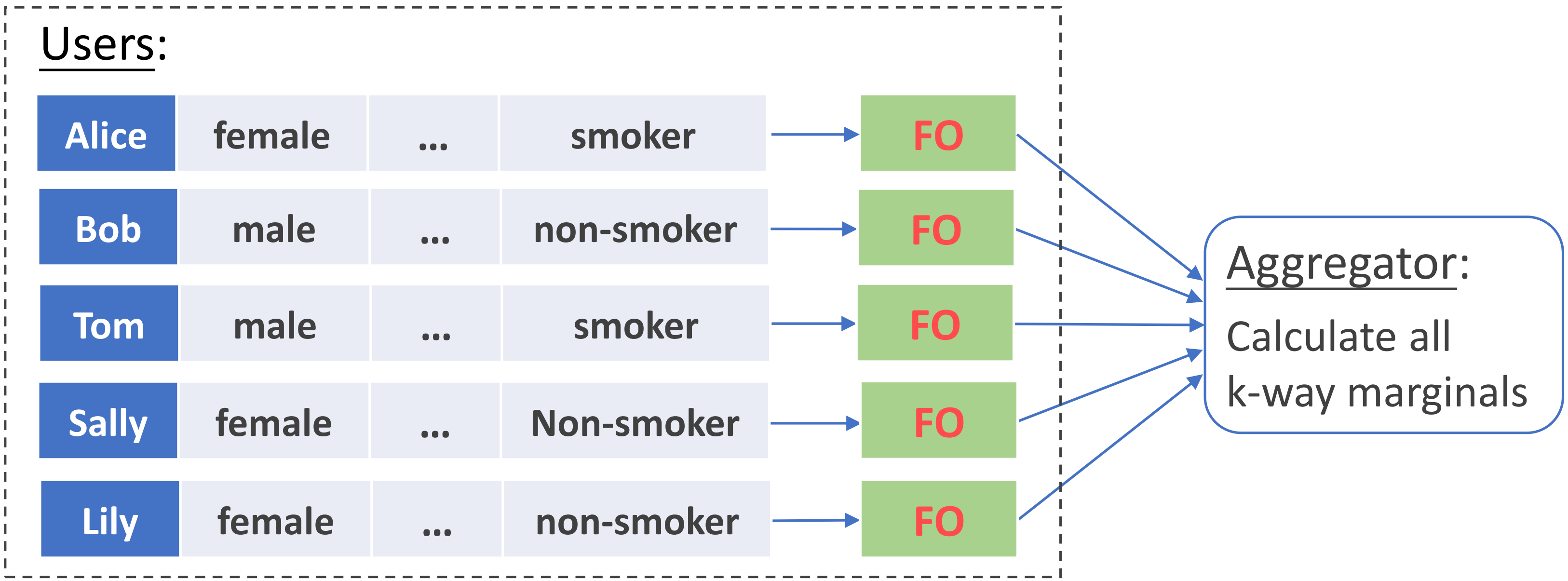
v	$F(v)$
< *, non-smoker >	0.55
< *, smoker >	0.45

1-way marginal

Marginal Release

- Each marginal is a frequency distribution, which can be seen as a frequency oracle problem

- Marginal release in local setting:



- Challenge: large number of attributes d

Marginal Release

- Straightforward method (1)

All attributes → Frequency Oracle → Full contingency table → All k-way marginals

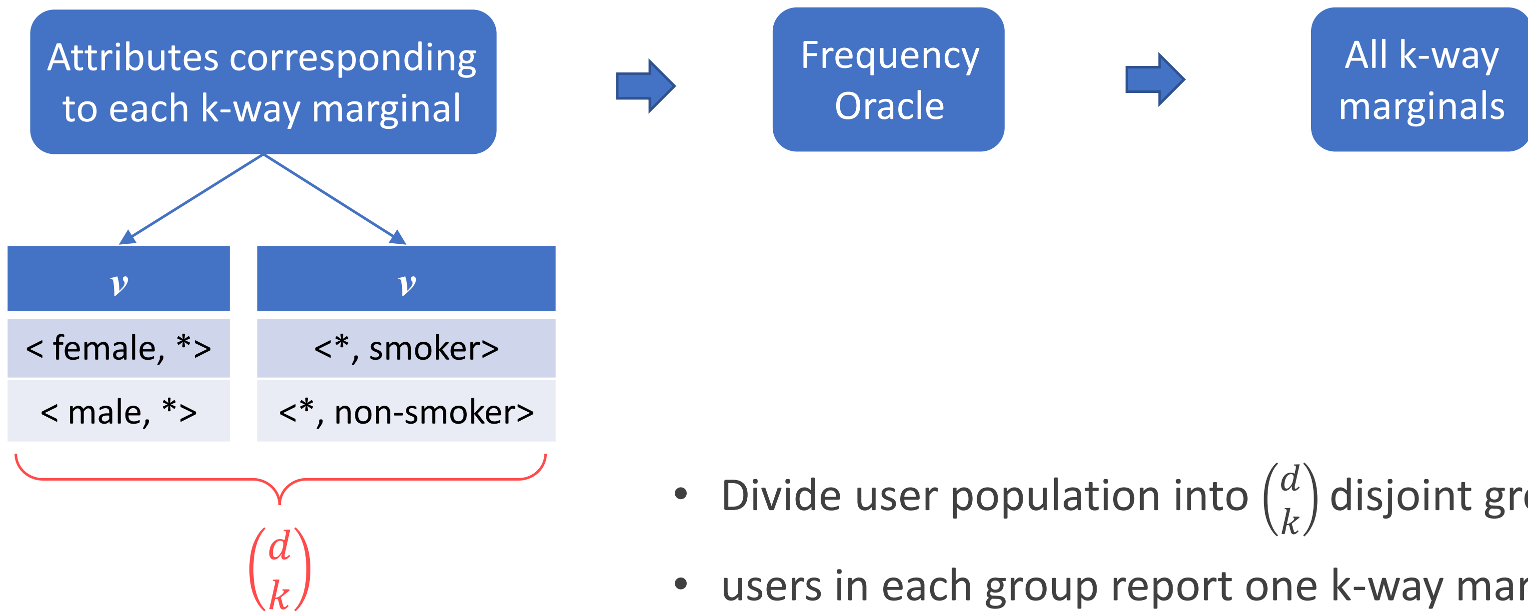
Gender	Smoke
female	smoker
male	non-smoker
male	smoker
…	…
female	non-smoker

v	$F(v)$
< female, non-smoker >	0.35
< female, smoker >	0.15
< male, non-smoker >	0.1
< male, smoker >	0.4

- Drawback:
  - Estimation error is exponential proportional to d, $Var = O(2^d)$
  - Time and space complexity are exponential proportional to d.

Marginal Release

- Straightforward method (2)



- Divide user population into $\binom{d}{k}$ disjoint groups
- users in each group report one k-way marginal

- Drawback:
  - When $\binom{d}{k}$ becomes large, each user contributes less information to each marginal
  - Still cause large estimation error, $Var = O(2^k \cdot \binom{d}{k})$

Marginal Release

- Fourier Transformation Method [SIGMOD' 18]

All attributes → Unary encoding → Fourier Transformation → Sample and randomize → All k-way marginals

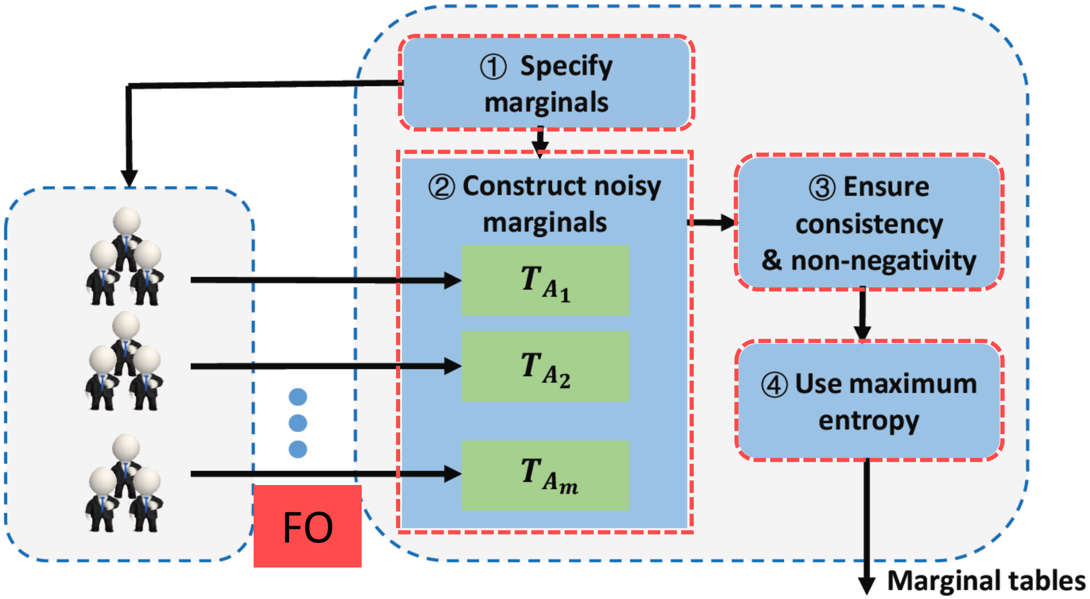
- Key observation:
  - Calculation of a k-way marginal requires only a few coefficients in the Fourier domain (values in marginals → Fourier coefficients)
  - Better than the two straightforward methods, in theory and in practice

$$Var = O\left(\sum_{s=0}^{k} \binom{d}{k}\right)$$

- Drawback:
  - To reconstruct all k-way marginals, there will be several coefficients to be estimated.
  - Perform poorly for large k

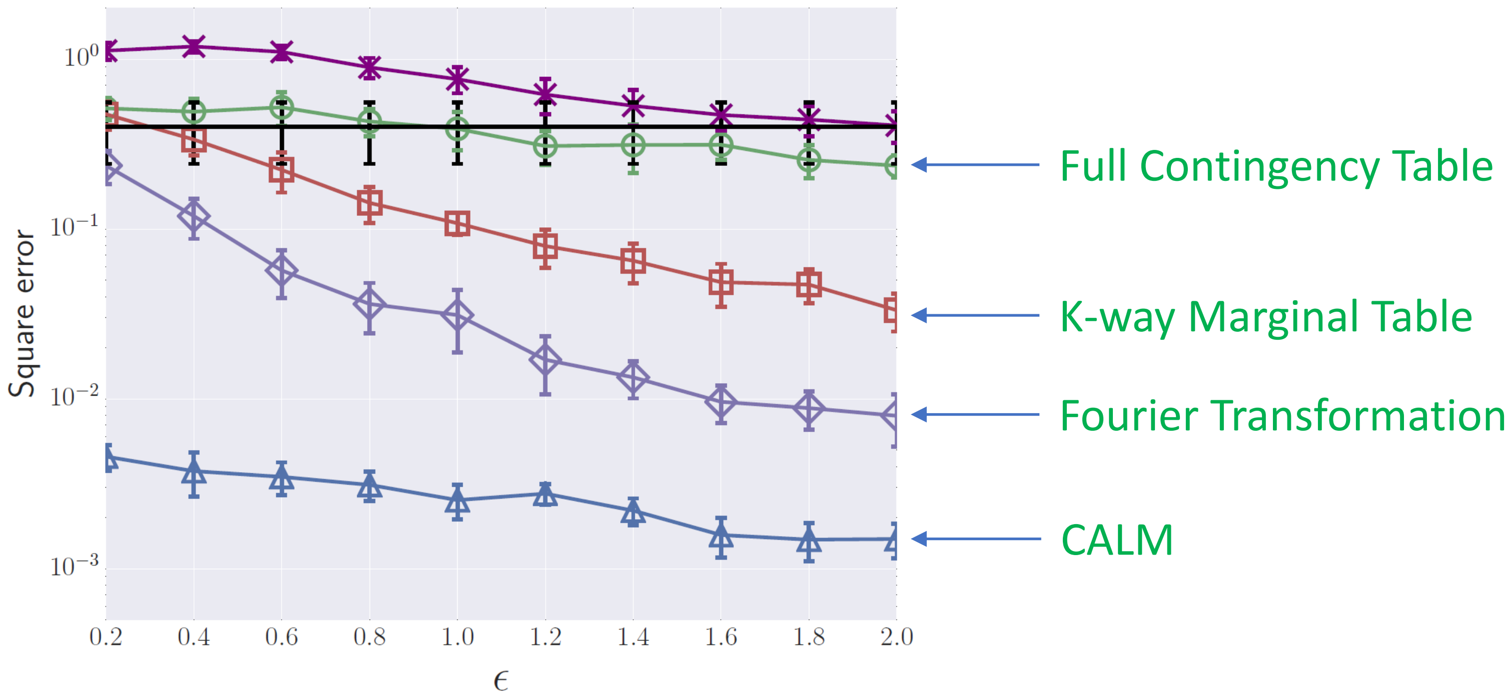
Marginal Release

- CALM: Consistent Adaptive Local Marginal [CCS’ 18]
- Intuition:
  - First construct a set of candidate marginals
  - Use the above marginals to reconstruct other unknown marginals



Marginal Release

- CALM: Consistent Adaptive Local Marginal [CCS' 18]



(b) $n = 2^{16}$, $d = 16$, $k = 3$

- The estimation error of CLAM decreases by 1-2 orders of magnitude.

Outline

- Current Research Problem
  - Marginal Release
  - **Graph Data Mining**
  - Key-Value Data Collection

- Open Problems and New Directions
  - Iterative Interaction
  - Privacy-Preserving Machine Learning
  - Theoretical underpinning

Graph Data Mining

- Graph data mining has numerous applications in web, social network, transportation and knowledge base.

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Adjacency bit vector

- Node-LDP: LDP definition applies to any two adjacency bit vectors

l_i :	1	1	1	1	1	0	0	0	0	0
l_j :	0	0	0	0	0	1	1	1	1	1

- Edge-LDP: LDP definition applies to any two adjacency bit vectors that only differ in one bit

l_i :	1	1	1	1	1	0	0	0	0	0
l_j :	1	1	1	1	1	1	0	0	0	0

- Results so far only for <u>edge-LDP</u> definition

Graph Data Mining

- Synthetic social graph generation [CCS' 17]
- Randomized Neighbor List (RNL)
  - Perturb each bit of the adjacency bit vector with RR
  - Retain some neighborhood information, but introduce a lot of fake edges

Facebook

4,039 nodes
88,234 edges

$\epsilon = 1$ →

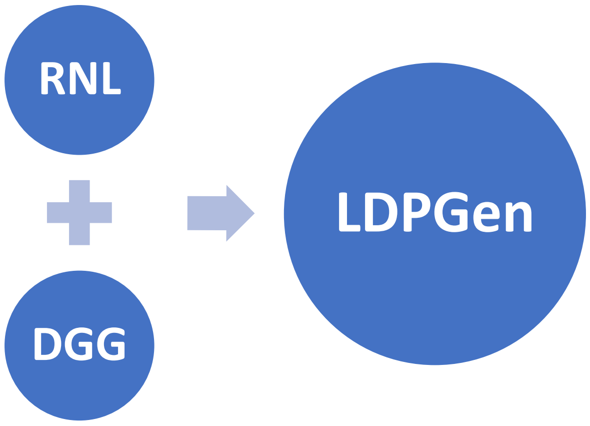
4,039 nodes
4,427,047 edges

98% fake edges

- Degree-based Graph Generation (DGG)
  - Perturb degree of each node with edge-LDP (Laplace noise)
  - Generate a synthetic graph by graph generation model (BTER)
  - Accurately collect statistics, but lose neighborhood information

Graph Data Mining

- RNL vs. DGG: neither baseline is very satisfying
- LDPGen: group-based graph generation
  - Strike a balance between noise and information loss
  - An iterative solution
  - Each user sends more information to aggregator
    (a single degree → a degree vector)



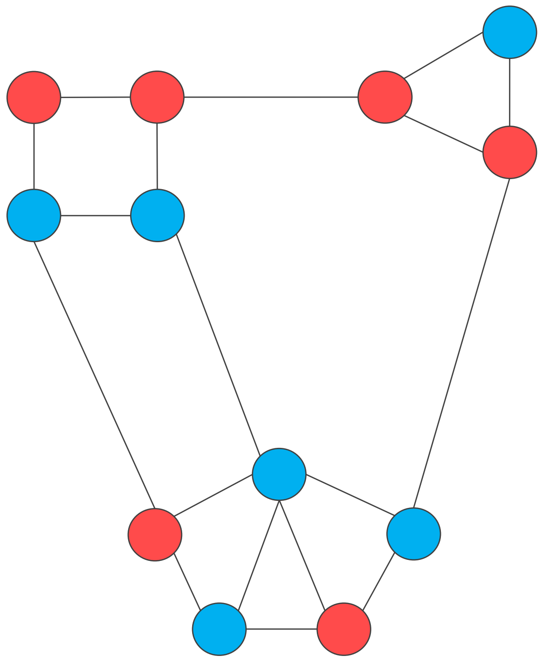
Graph Data Mining

- Three phases of LDPGen

1. Initial grouping: aggregator randomly partitions users into k groups

   - Users report <u>noisy degree vector</u> of their links to these groups

   - Aggregator optimizes k and refines grouping



k = 2

Graph Data Mining

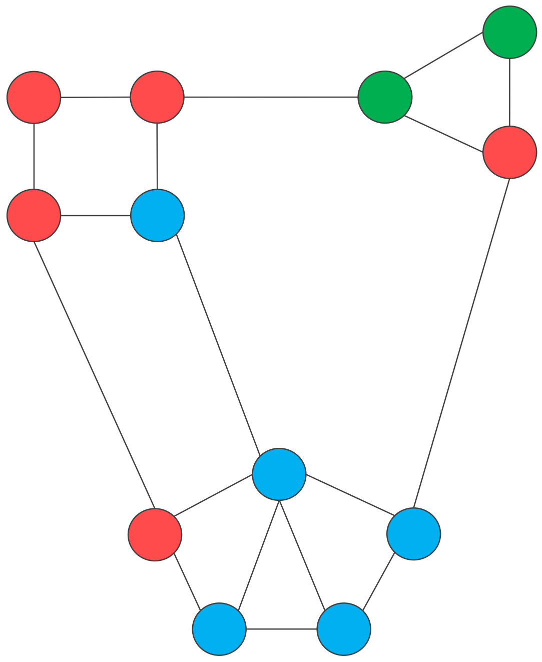
- Three phases of LDPGen

1. Initial grouping: aggregator randomly partitions users into k groups

    - Users report noisy degree vector of their links to these groups

    - Aggregator optimize k and refine grouping

2. Grouping refinement: aggregator partitions users with similar degree distribution into new groups



k = 3

Graph Data Mining

- Three phases of LDPGen

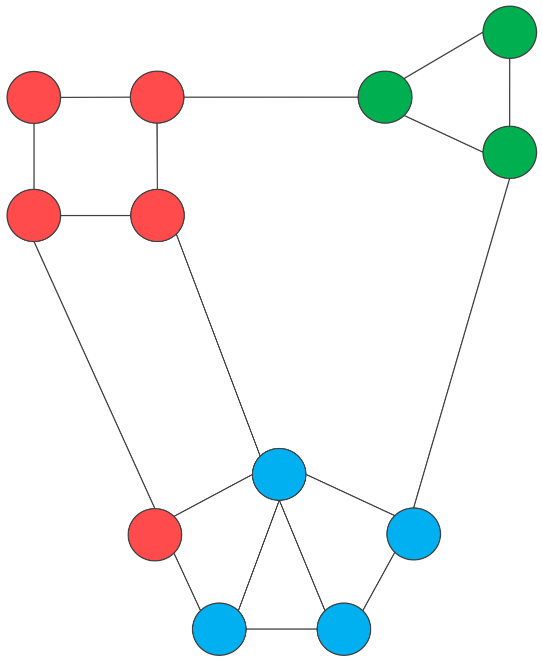
1. Initial grouping: aggregator randomly partitions users into k groups

   - Users report noisy degree vectors of their links to these groups

   - Aggregator optimize k and refine grouping

2. Grouping refinement: aggregator partitions users with similar degree distribution into new groups

   - Users report again noisy degree vectors of their links to the new groups



k = 3

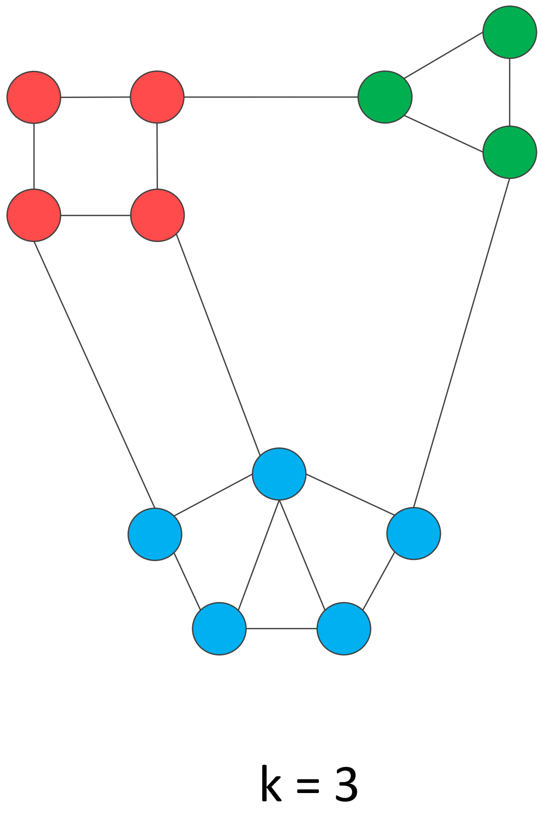
Graph Data Mining

- Three phases of LDPGen

1. Initial grouping: aggregator randomly partitions users into k groups
   - Users report noisy degree vector of their links to these groups
   - Aggregator optimize k and refine grouping

2. Grouping refinement: aggregator partitions users with similar degree distribution into new groups
   - Users report again noisy degree vectors of their links to the new groups

3. Graph generation: sample a corresponding graph from BTER model



Outline

- Current Research Problem
  - Marginal Release
  - Graph Data Mining
  - **Key-Value Data Collection**
- Open Problems and New Directions
  - Iterative Interaction
  - Privacy-Preserving Machine Learning
  - Theoretical underpinning

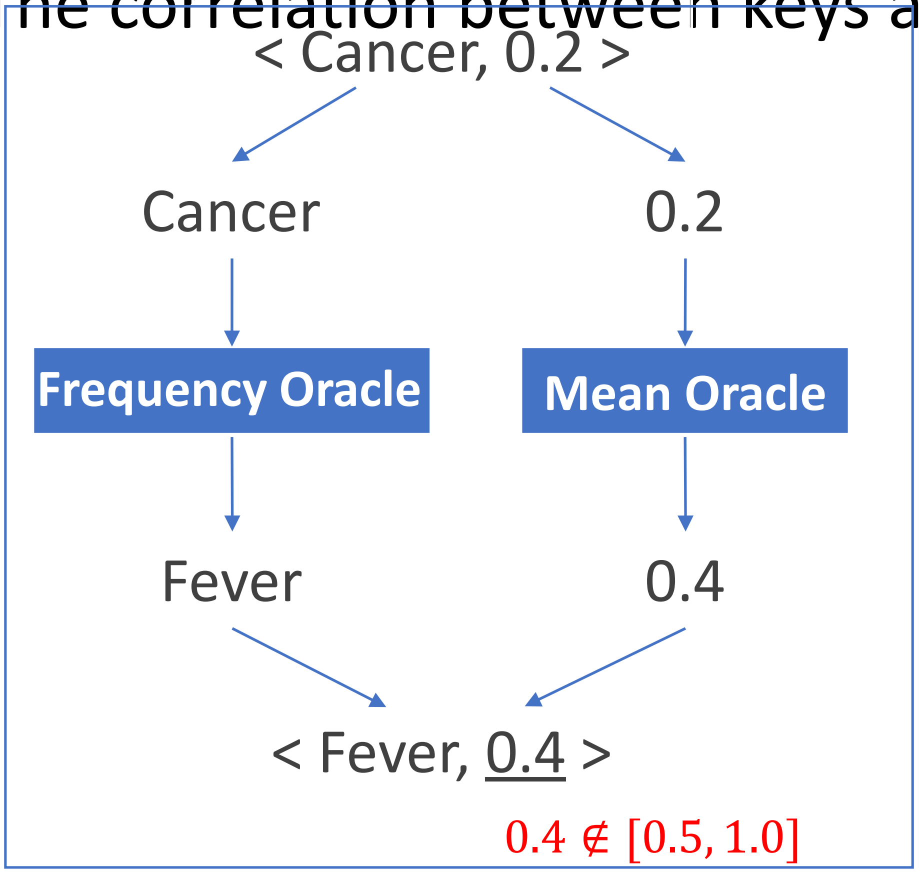
Key-Value Data Collection

- Key-value pair is an popular data model $< Key, Value >$

- To estimate the average … ach app

Key-Value Data Collection

- The correlation between keys and values



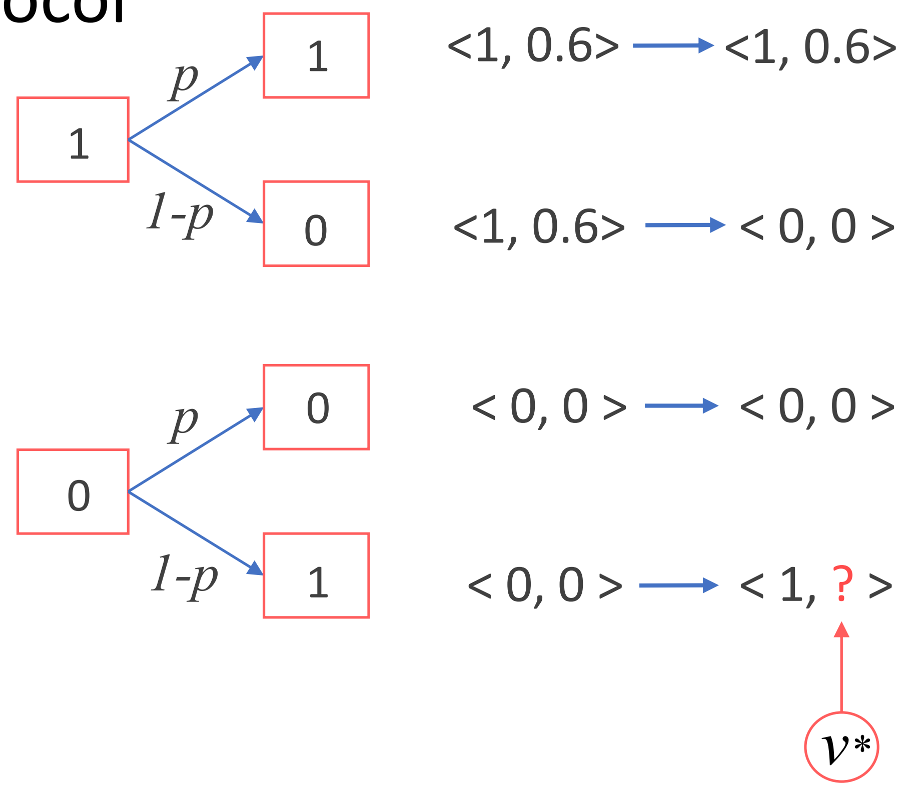
Disease	Domain
Cancer	[0, 0.35]
HIV	[0.3, 0.6]
Fever	[0.5, 1.0]

Key-Value Data Collection

- PrivKV: iterative model [S&P’ 19]

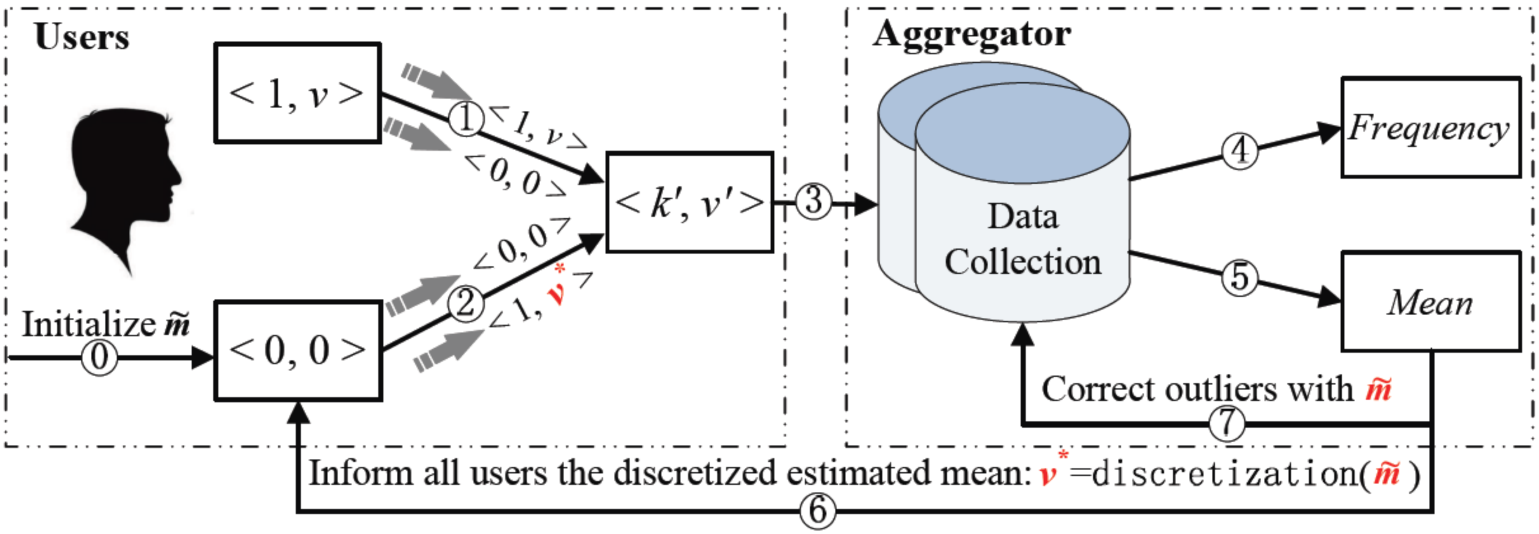
- Perturbation protocol

Users	Item
Alice	< 0, 0 >
Bob	< 1, 0.6 >
Chris	< 0, 0 >
Tom	< 1, 0.8 >



Key-Value Data Collection

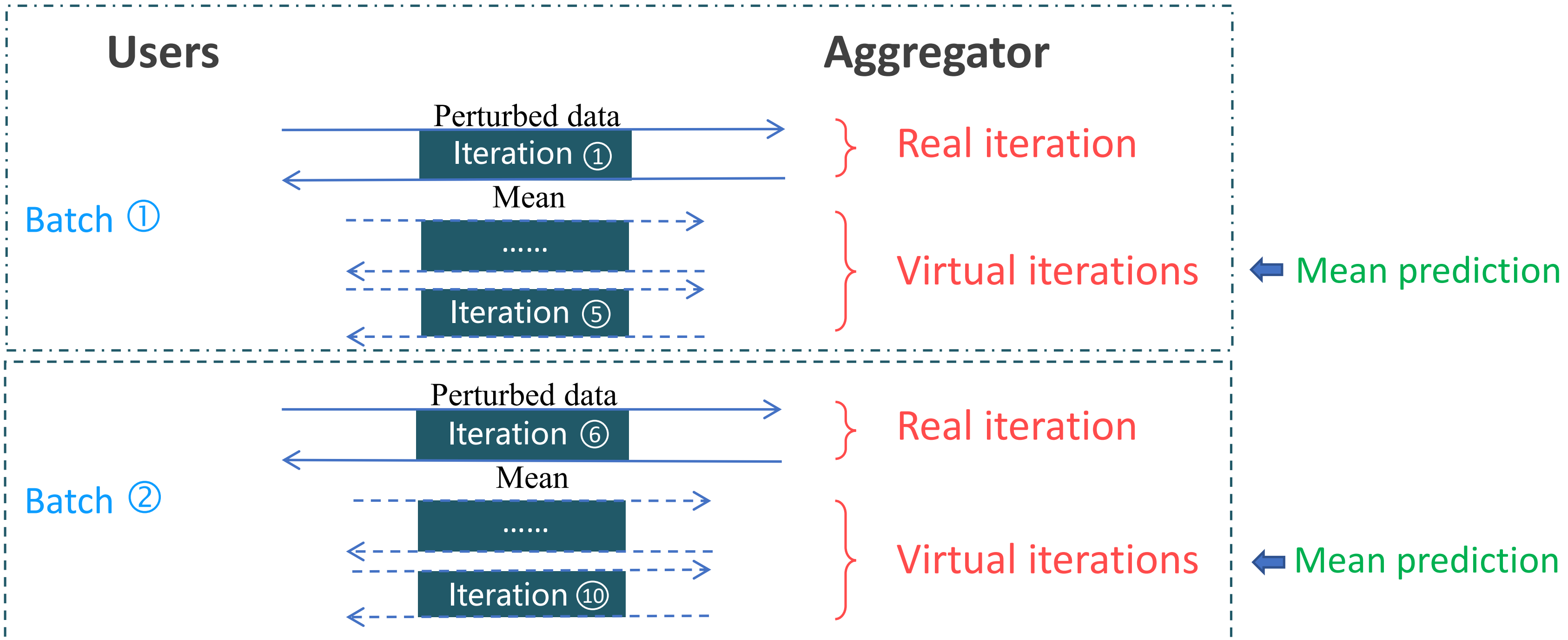
- Iterative model



- Analysis
  - High accuracy: the estimated mean gradually approaches the ground truth.
  - High communication bandwidth with multiple iterations

Key-Value Data Collection

- Batch processing and virtual iterations



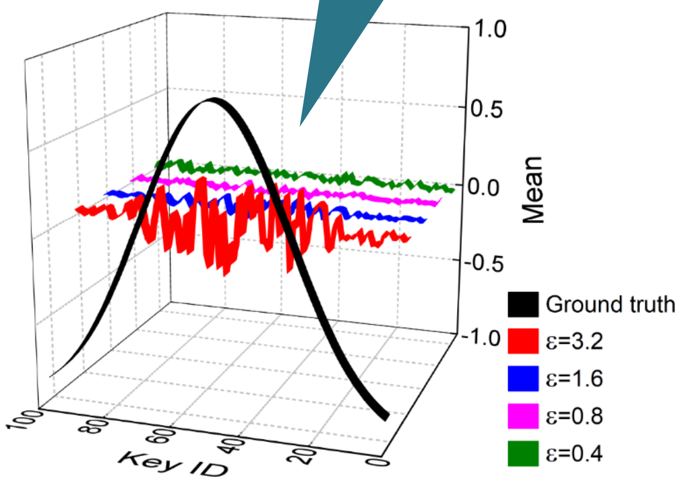
- Analysis
  - Without user involvement in virtual iterations —reduce network transmission overhead
  - No privacy budget cost in virtual iterations — improve accuracy

Key-Value Data Collection

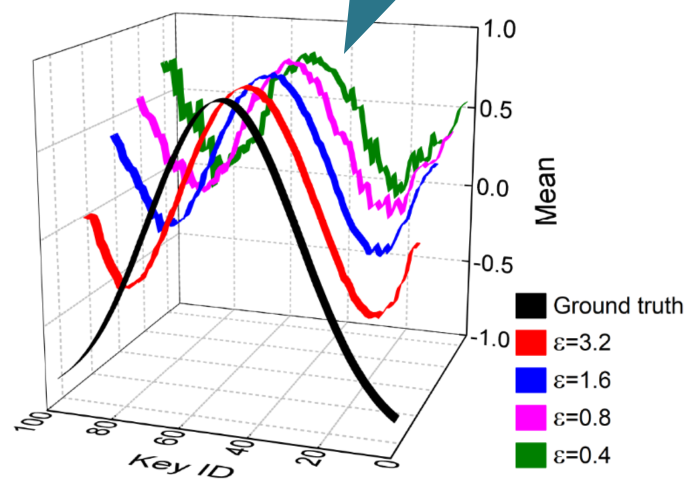
- Key-value correlation

Deviate from the true distribution

similar distribution as the real mean.



$k\text{-}RR\&Harmony$



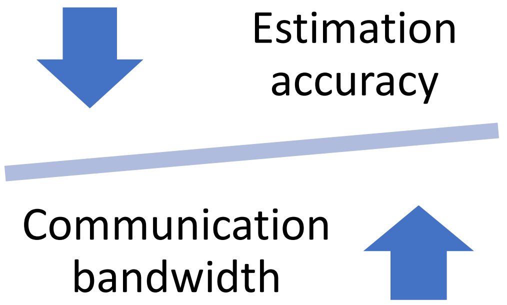
$PrivKVM$

Outline

- Current Research Problem
  - Marginal Release
  - Graph Data Mining
  - Key-Value Data Collection
- Open Problems and New Directions
  - Iterative Interaction
  - Privacy-Preserving Machine Learning
  - Theoretical underpinning

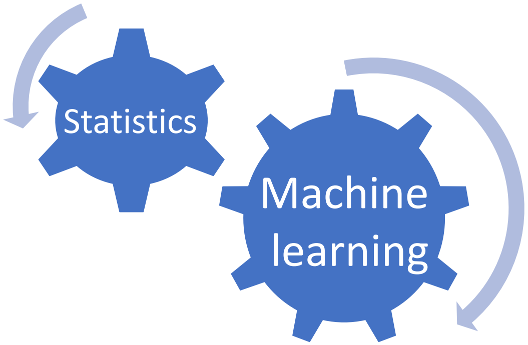
Iterative Interactions

- Access the original data multiple times
  → multiple rounds of interactions
- In each round, the aggregator poses new queries in the light of previous response
- Existing works:
  - Heavy hitter estimation [CCS’ 16]
  - Synthetic graph generation [CCS’ 17]
  - Key-value data collection [S&P’ 19]
  - Machine learning model [ICDE’ 19]
- The effectiveness of iterations ?



Privacy-Preserving Machine Learning

- Machine learning needs to learn from real data
  - LDP incurs heavy perturbation
- Traditional machine learning assumes centralized data
  - Each user only has a local view under LDP
- Existing works:
  - Simple machine learning models, e.g., linear regression, logistic regression and support vector machine [ICDE' 19]
  - Single-round machine learning [S&P' 17] [ICML' 17]
- Machine learning with LDP ?



Theoretical Underpinnings

- LDP emerged most recently from the theory literature
  - What can we learning privately? [FOCS' 08]
  - Local privacy and statistical minimax rates [FOCS' 13]

- Still many theoretical questions about LDP
  - What are the lower bounds of the accuracy guarantee?
  - Is there any benefit from adding an additive "relaxation" δ to the privacy definition?

$$\Pr[A(s) = s^*] \leq e^{\varepsilon} \cdot \Pr[A(s') = s^*] + \delta$$

  - How to minimize the amount of data collected from each user to a single bit?

Conclusions

- Privacy-preserving data release is an important and challenging problem.
- Local Differential Privacy is a promising privacy model and has been widely adopted.
- Lots of current research that can be applied to mobile
  - Histogram estimation, frequent itemset mining
  - marginal release, graph data mining
  - key-value data collection, private spatial data aggregation
- Lots of opportunity for new work:
  - Optimal mechanisms for local differential privacy
  - High-dimensional data perturbation protocol
  - Unstructured data: text, image, video

Thank you!