Dependence Makes You Vulnerable: Differential Privacy Under Dependent Tuples

Changchang Liu[1], Supriyo Chakraborty[2], Prateek Mittal[1]
Email: [1]{cl12, pmittal}@princeton.edu, [2]supriyo@us.ibm.com,
[1] Princeton University, [2]IBM T.J. Watson Research Center

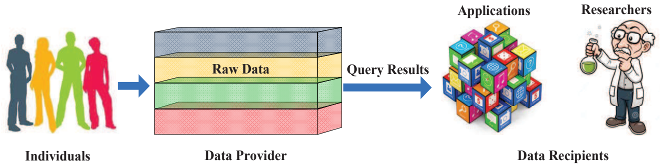
February 23, 2016

Data Privacy

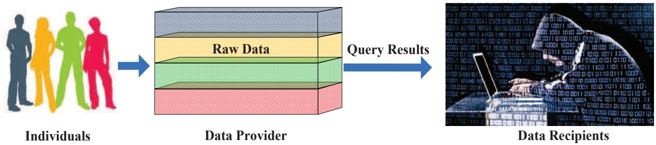
- Privacy is important!
  - Snowden case
  - G20 summit breach
  - iCloud photo breach

    ...

Direct Release Data Would Compromise Privacy!



Direct Release Data Would Compromise Privacy!



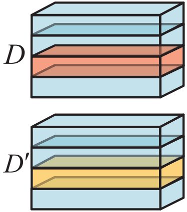
Obfuscate Data before Release to Protect Privacy

Existing Privacy Metrics

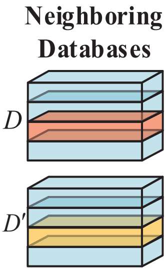
- Differential Privacy [ICALP '06]
- Pufferfish Privacy [PODS '12]
- Membership Privacy [CCS '13]
- Blowfish Privacy [SIGMOD '14]

ε-Differential Privacy (DP)

Neighboring Databases



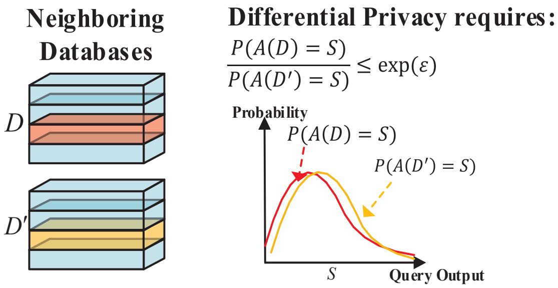
ε-Differential Privacy (DP)



Differential Privacy requires:

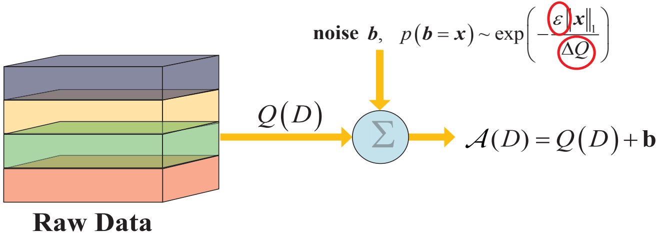
$$\frac{P(A(D) = S)}{P(A(D') = S)} \leq \exp(\varepsilon)$$

ε-Differential Privacy (DP)



The adversary's ability to infer the individual's information is bounded!

Laplace Perturbation Mechanism



ε is the privacy budget

Q is the query function

ΔQ is the global sensitivity of Q: $max_{D,D'} \|Q(D) - Q(D')\|_1$

Limitations for Differential Privacy (DP) Mechanisms

Implicitly assume **independent** tuples

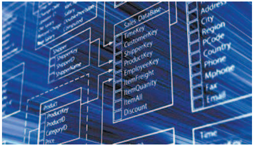
Limitations for Differential Privacy (DP) Mechanisms

In reality, however, tuples are **correlated**

- large volume
- rich semantics
- complex structure

Data correlation exists almost everywhere

(a) social network data



(b) business data

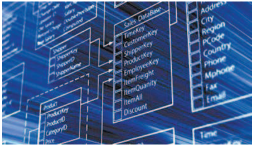
(c) mobility data

(d) medical data

Data correlation exists almost everywhere

friendships
interactions

(a) social network data



(b) business data

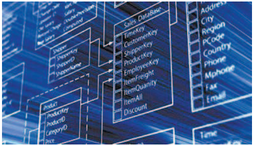
(c) mobility data

(d) medical data

Data correlation exists almost everywhere

friendships
interactions

(a) social network data



(b) business data

financial
transactions

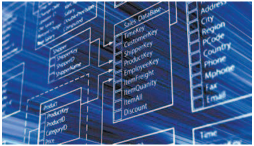
(c) mobility data

(d) medical data

Data correlation exists almost everywhere

friendships
interactions

(a) social network data



(b) business data

financial
transactions

communication
records

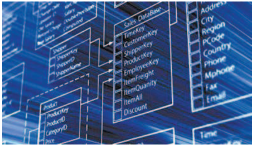
(c) mobility data

(d) medical data

Data correlation exists almost everywhere

friendships interactions

(a) social network data



(b) business data

financial transactions

communication records

(c) mobility data

(d) medical data

disease transmission

Our Objective

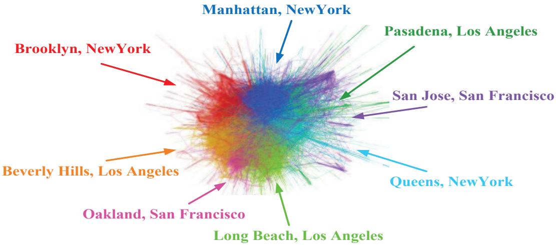
Incorporate correlated data in differential privacy

Differential Privacy under Dependent Data

- Inference Attack for DP based on Correlated Tuples
- Dependent Differential Privacy (DDP)
- Experimental Results

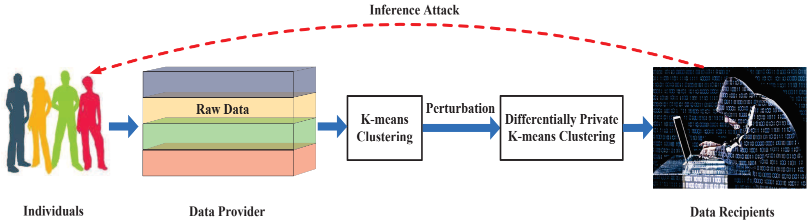
Correlation in Gowalla Location Dataset

Gowalla location dataset: $6,969$ users, $98,802$ location records
Gowalla social dataset: $6,969$ users, $47,502$ edges

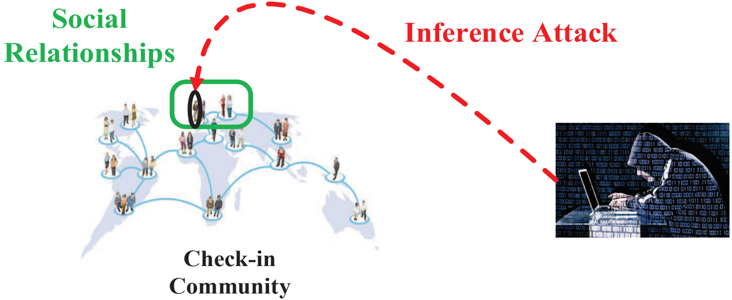


Inference Attack on DP via K-Means Query

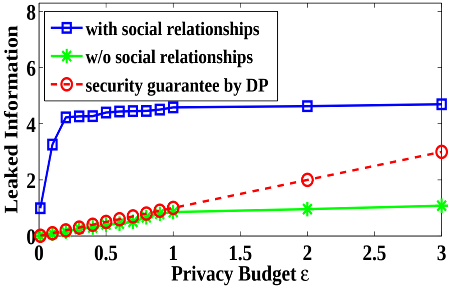
Differentially Private K-means for Gowalla Location Dataset



Inference Attack

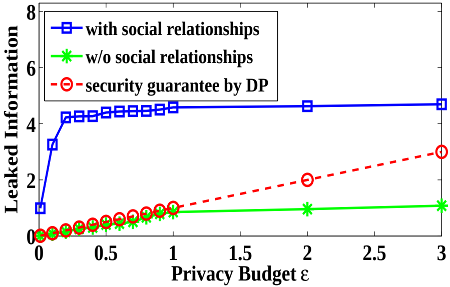


Inference results by using correlation



Exploiting correlation, one can infer more information!

Inference results by using correlation

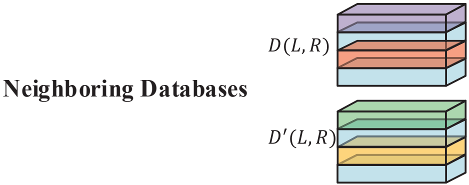


Exploiting correlation, one can infer more information!
Exploiting correlation can break DP security guarantees!

Differential Privacy under Dependent Data

- Inference Attack for DP based on Correlated Tuples
- Dependent Differential Privacy (DDP)
- Experimental Results

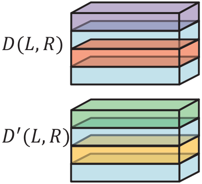
ε-Dependent Differential Privacy (DDP)



- R is probabilistic dependence relationship among the L dependent tuples

ε-Dependent Differential Privacy (DDP)

Neighboring Databases

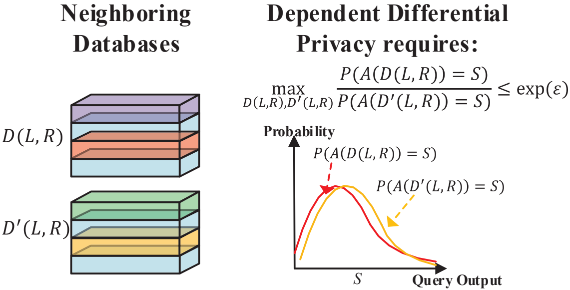


Dependent Differential Privacy requires:

$$\max_{D(L,R),D'(L,R)} \frac{P(A(D(L,R)) = S)}{P(A(D'(L,R)) = S)} \leq \exp(\varepsilon)$$

- R is probabilistic dependence relationship among the L dependent tuples

ε-Dependent Differential Privacy (DDP)



- R is probabilistic dependence relationship among the L dependent tuples
- The adversary's ability to infer the individual's information is bounded even if the adversary has access to data correlation R.

Dependent Perturbation Mechanism

- Augment conventional LPM with additional noise relevant to ρ_{ij}
- **Dependent coefficient** ρ_{ij}
  - extent of dependence of D_j on the modification of D_i

Dependent Coefficient

Laplace noise in dependent perturbation mechanism

$$\exp\left\{-\frac{\varepsilon}{\text{Sensitivity}_i + \rho_{ij} \times \text{Sensitivity}_j}\right\}$$

Dependent coefficient satisfies: $0 \leq \rho_{ij} \leq 1$

- $\rho_{ij} = 0$: standard differential privacy (independent setting)
- $\rho_{ij} = 1$: fully dependent setting
- ρ_{ij}: formulate correlation from privacy perspective

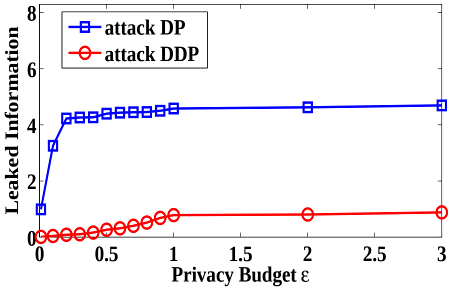
Limitations of Dependent Coefficient

The exact computation of ρ_{ij}

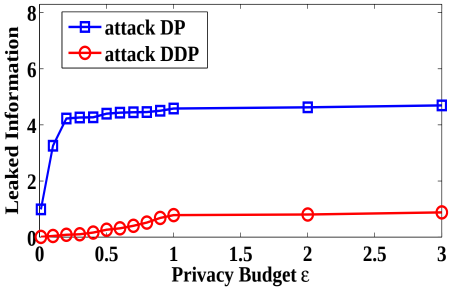
relies on knowledge of data generation model

Differential Privacy under Dependent Data
Inference Attack for DP based on Correlated Tuples
Dependent Differential Privacy (DDP)
Experimental Results

Resilience to Inference Attack



Resilience to Inference Attack



DDP is more resilient to inference attack

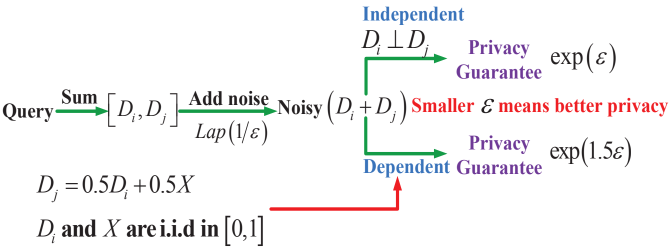
Further Analysis and Experiments

- Composition Property
  - Sequential/parallel composition property
- Theoretical utility analysis
- Different classes of queries
  - Machine learning queries
  - Graph queries

Conclusion and Future work

- Incorporate correlation into differential privacy
  - Dependent differential privacy
  - More resilient to inference attack
- Alternative data generation models in the future work

Appendix1: Dependence between tuples can seriously degrade the privacy guarantees provided by the existing DP mechanisms



Appendix 2: Model to Compute Dependent Coefficient

Here, we consider to utilize the friend-based model to compute the probabilistic dependence relationship, where a user's location can be estimated by her friend's location based on the distance between their locations. Specifically, the probability of a user j locating at $\mathbf{d}_j$ when her friend i is locating at $\mathbf{d}_i$ is

$$P(D_j = \mathbf{d}_j | D_i = \mathbf{d}_i) = a(\|\mathbf{d}_j - \mathbf{d}_i\|_1 + b)^{-c} \tag{1}$$

where $a > 0, b > 0, c > 0$.