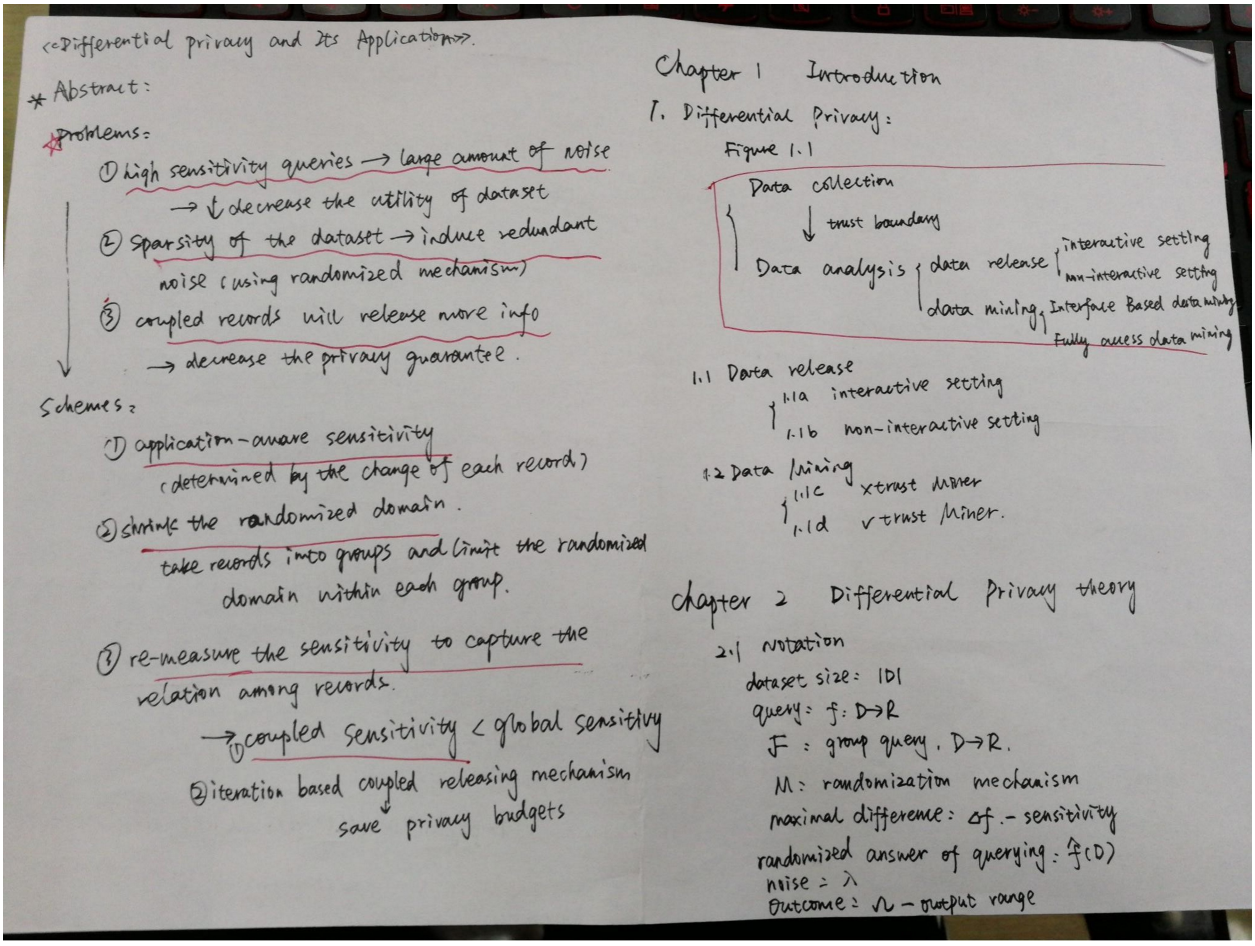
《differential privacy and its application》

Sunday, October 20, 2019 11:46 PM



《Differential privacy and its Application》.

* Abstract:

Problems:

① high sensitivity queries → large amount of noise → ↓ decrease the utility of dataset

② Sparsity of the dataset → induce redundant noise (using randomized mechanism)

③ coupled records will release more info → decrease the privacy guarantee.

Schemes:

① application-aware sensitivity (determined by the change of each record)

② shrink the randomized domain. take records into groups and limit the randomized domain within each group.

③ re-measure the sensitivity to capture the relation among records.

→ ① coupled sensitivity < global sensitivity

② iteration based coupled releasing mechanism → save privacy budgets

Chapter 1 Introduction

1. Differential Privacy:

Figure 1.1

Data collection
↓ trust boundary
Data analysis { data release { interactive setting / non-interactive setting
data mining { Interface Based data mining / Fully access data mining

1.1 Data release
{ 1.1a interactive setting
1.1b non-interactive setting

1.2 Data mining
{ 1.1c ×trust miner
1.1d √trust miner.

Chapter 2 Differential Privacy theory

2.1 notation

dataset size: $|D|$
query: $f: D \to R$
$\mathcal{F}$: group query. $D \to R$.
M: randomization mechanism
maximal difference: Δf - sensitivity
randomized answer of querying: $\hat{f}(D)$
noise: λ
Outcome: $\mathcal{R}$ - output range

Definition 1 $(\varepsilon-\tau)$-differential privacy

$Pr[M(D)\in\Omega] \le e^{\varepsilon}\cdot Pr[M(D')\in\Omega]+\tau$

ε: privacy budget

τ: violation probability

Definition 2: Sequential Composition

$M=\{M_1,\cdots M_m\}$ sequential performed on the dataset,

the M will provide $(m\cdot\varepsilon)$-differential privacy

↑ each M_i provides ε privacy guarantee.

Definition 3: Parallel Composition

$\max\{\varepsilon_1,\cdots\varepsilon_m\}$-differential privacy

2.2 The sensitivity { global sensitivity / local sensitivity

Definition 1: Global sensitivity

$$\Delta f_{GS} = \max_{D,D':d(D,D')=1} \|f(D)-f(D')\|_1.$$

↓ only works well with low sensitivity

Definition 2: Local sensitivity

$$\Delta f_{LS} = \max_{?'} \|f(D)-f(D')\|_1.$$

$\therefore \Delta f_{GS} = \max_{\delta} f_{LS}.$

generates less noise for queries with high sensitivity

2.3 Mechanism { Laplace: numeric queries / Exponential: non-numeric

- Laplace

probability: $p(x)=\frac{1}{2b}e^{-\frac{|x|}{b}}$

Definition 1: Laplace mechanism.

$M(D)=f(D)+Lap(\frac{\Delta f}{\varepsilon})$.

Definition 2: Exponential mechanism.

score function: $q(D,\psi)$, represent how good an output ψ is for D.

$M(D)=$ { return r with the probability $\propto \exp(\frac{\varepsilon q(D,r)}{2\Delta q})$ }

2.4 Differential Private Data Release

{ interactive setting / non-interactive setting: improve the accuracy of query results

↓ research lines

① query release. { batch queries release / group-based release

② group-based tech in data release

③ contigency table release

④ release a synthetic dataset

2.4.1 Interactive Data Release

query f_i will only be answered after the answer of previous query f_{i-1} released.

$$Pr_{f\in F}[|f(D)-\underline{M(f(D))}|\le\alpha]\ge 1-\delta$$

perturbation on query result.

2.4.2 mechanism design for Interactive Data Release

{ Naive Laplace / Query Separation / Iteration.

① naive Laplace: most popular

cons { maximal number(query) is limited in sub-linear n / noise magnitude will be large when dealing with continuous attribute

② Query separation:

Median mechanism, separating queries into hard and easy ones

pros: answer exponentially more queries than Naive Laplace

cons: time complexity

③ Iteration:

1) Private Multiple Weights. (PMW)

pros: saved the privacy budget & decrease noise.

2) Iterative Database Construct. (IDC)

when a significant difference occurred, the mechanism updates the current dataset for next iteration.

pros: more general than PMW.

2.4.3. Histogram Release.

Definition 1: Histogram Representation.

Laplace: cons - large noise

① partition the attributes into several mutual groups.

② kd-tree based partition method to generate nearly uniform partitions

↓ DPCube

③ optimal partition method by minimizing the sum of Squared Error (SSE)

1) Noise first

2) structure first

④ maintain the consistency:

constrained inference { ① Sorted constraints / ② hierarchical constraints.

2.4.4 non-Interactive Data Release.

{ pros: more queries, higher flexibility

cons: too much noise, ↓utility

↓ batch queries release

research lines { contigency table release. / group-based release. / Sanitized dataset release

① batch queries release

② group-base release.

③ contigency table release

④ synthetic Dataset Release.

{ how many records → α-accurate result / algorithm → obtain the synthetic dataset in polynomial time.

Definition 1: $(\alpha-\delta)$-usefulness.

$pr\{|f(M(D))-f(D)|\le\alpha\}\ge 1-\delta.$

2.5 Differentially Private Data Mining

{ data mining with interface - / data mining with full access -

2.5.1. data mining with interface.

interface { sub-linear Queries (SuLQ) / Privacy Integrated Queries platform (PINQ)

① SuLQ: $\sum_i f(r_i)+N(0,R)$

② PINQ: Laplace, Exponential

2.5.2 classification

① SuLQ framework – ID3 algorithms.

cons: { ① privacy budget arrangement
② dealing with continuous attributes

② PINQ improve { 1) Exponential mechanism in the attributes choosing step
2) divided into ranges by every possible splitting value

2.5.3 clustering

① SuLQ + K-means clustering

$\bar{S}_j = \begin{cases} \text{SuLQ}(f(r_i)) := 1 & \text{if } j = \arg\min_j \\ 0 & \text{otherwise} \end{cases}$

$\bar{C}_j = \begin{cases} \text{SuLQ}(f(r_i)) := r_i, & \text{if } j = \arg\min_j \\ 0 & \text{otherwise} \end{cases}$

↓ noise

② Local sensitivity on K-means clustering

2.5.4 Data Mining with Full access.

① ID3
② logistic regression { 1) inject Laplace noise in classifier ✗
2) add noise to the loss function (LCD). ✓

2.5.5 Frequent Itemset Mining

to discover itemsets that frequently appear in dataset

① top-K frequent itemsets mining algorithm DiffFIM
{ private selection : truncated frequency
frequency perturbation

② PrivBasis: ensure high accuracy

2.6 Applications of Differential privacy

applications: recommender system,
tagging recommender system
coupled differential privacy

2.6.1 privacy preserving recommender system

CF: collaborative Filtering { neighborhood-based methods
model-based methods

? privacy leak risk, kNN attack

2.6.2 Privacy Preserving Tagging recommender system.

① tag suppression approach.

model the user's profile using a tagging ~~system~~ histogram and eliminate sensitive tags from this profile.

cons: only release an incomplete dataset.

2.6.3 Differential privacy for a coupled dataset.

Chapter 3 Differentially Private Neighborhood-based Collaborative Filtering 协同过滤.

3.1 Introduction.

application-aware sensitivity in the context of neighborhood-based collaborative filtering

① Collaborative filtering (CF) { neighborhood-based methods
model-based methods.

↓

② PNCF: Private Neighbor Collaborative Filtering

{ 1) all possible privacy leakage should be considered.
2) sensitivity and mechanism

3 research lines { 1) How to preserve neighborhood privacy – Private Neighbor selection
2) How to define sensitivity for recommendation purpose – recommendation-aware sensitivity
3) How to design the exponential mechanism for CF?

3.2 Fundamentals of a recommender System.

D can be represented by column vectors:

$D = [\vec{t}_1, \vec{t}_2, \cdots \vec{t}_m]$, and $\vec{t}_i = [r_{1i}, r_{2i}, \cdots r_{ni}]^T$

$S(i,j)$ denotes the similarity between t_i with t_j

① Collaborative Filtering (CF)

neighborhood-based methods { neighbor selection
rating prediction

Similarity: { Pearson correlation coefficient (pcc) / cosine-based similarity (cos) }

a. Pearson correlation coefficient:

$$sim(i,j) = \frac{\sum_{x \in U_{ij}} (r_{xi} - \bar{r}_i)(r_{xj} - \bar{r}_j)}{\sqrt{\sum_{x \in U_{ij}} (r_{xi} - \bar{r}_i)^2} \sqrt{\sum_{x \in U_{ij}} (r_{xj} - \bar{r}_j)^2}}$$

$\bar{r}$ is the average rating given by relative users.

b. cosine-based similarity

$$sim(i,j) = \frac{r_i \cdot r_j}{\|r_i\|_2 \, \|r_j\|_2}$$

Rating prediction stage:

item-based:

$$\hat{r}_{ai} = \frac{\sum_{j \in J_a} s(i,j) \cdot r_{a,j}}{\sum_{j \in J_a} |s(i,j)|}$$

user-based:

$$r_{ui} = \bar{r}_v + \frac{\sum_{u \in U_b} S_{vu}(r_{ux} - \bar{r}_u)}{\sum_u |S_{vu}|}$$

$\bar{r}_v$, $\bar{r}_u$ is the average rating given by user u and v.

S_{vu} denotes the similarity between v to u.

② knn attack to cF., $O(\log n)$.

③ related work

- Privacy Preserving Recommender System.
- Differential Private Recommender System

3.3 private Neighbor Collaborative Filtering

① overview of algorithm. (PNCF)

PNCF { private neighbor selection: prevent inferring neighbor / perturbation: prevent inferring rating }

② The private neighbor selection means privately select k neighbors from a list of candidates for the privacy preserving purpose.

A. Recommendation-Aware sensitivity, $RS(i,j)$.

B. private neighbor selection ++ accuracy

truncated similarity:

$$\hat{S}(i,j) = \max(s(i,j), S_k(i,\cdot) - w)$$

ensures no item in $N_k(t_i)$ has a similarity less than $S_k(i,\cdot) - w$. and is greater than $(S_k(i,\cdot) + w)$

3.4 Privacy and utility analysis.

① Privacy analysis.

$$S_{noise}(i,j) = S(i,j) + Lap\left(\frac{2 \cdot RS(i,j)}{\varepsilon}\right).$$

② Utility analysis.

3.5 Experiment and analysis.

research lines { ① How does PNCF perform on the privacy preserving issue from the perspective of neighbors? / ② How does the proposed recommendation-Aware sensitivity satisfy the recommendation purpose? / ③ How does PNCF perform compared to other-related methods? / ④ How does the privacy budget affect the performance of PNCF? }

① Datasets and measurements

ex: Netflix, Movielens

Mean absolute Error (MAE):

$$MAE = \frac{1}{|T|} \sum_{a,i \in T} |r_{ai} - \hat{r}_{ai}|$$

② Performance of PNCF.

③ performance of Recommendation-Aware sensitivity

④ performance of the private neighbor selection mechanism

⑤ Effect of Privacy Budget to determine privacy preserving level

chapter 4 Differentially private Tagging Recommender System.

research challenges { ① ignore the relationship between items/sources / ② randomized mechanism will result in a large magnitude of noise with million tags }

↓

schemes { ① generate a synthetic dataset which retains the relationship / ② shrink the randomized domain }

contribution: { ① utility ✓. PriTop. / ② model-based method, & shrink the randomized domain. / ③ privacy — utility. }

4.1 Fundamental of Tagging Recommender System.

① Notation:

$T(u_a, r_b)$: represent all tags, flagged by the user u_a, on resource r_b.

$P(u_a)$: a user u_a's profile.

$W(u_a)$: tagging's weights

② Tagging recommender System

4.2 private Tagging Release.

Private Topic-based Tagging Release (PriTop)

① overview of PriTop algorithm

1) Private Topic Model Generation: mask the topic distribution on tags

PriTM operations

2) Topic weight perturbation: mask the weight of tags

3) Private tag selection:

(2) Private Topic Model Generation

creates multiple topics according to the resources & tags

topic models: Latent Dirichlet Allocation (LDA)

3 steps:
1) LDA Model construction: Generating a LDA model using the Gibbs Sampling approach
2) Private Model Generation: Adding Laplace noise to the LDA model to create a private model
3) Topic-based profile Generation: Creating a user's profile according to the private LDA model

A. LDA model construction

B. Private Model Generation

C. Topic-based profile Generation

(3) Topic weight Perturbation

$$\hat{w}_z(u_a) = w_z(u_a) + Laplace\left(\frac{4}{\varepsilon}\right)^K$$

(4) Private tag selection: ~~retag~~ replace original tags with selected new tags

} retain the utility of tags
mask the similarities between tags

4.3 Privacy and utility analysis

4.4 Experiment & Analysis.

Chapter 5 coupled differential privacy for Non-IID datasets

challenges:
1) how to identify and represent coupled records
2) coupled records will be large
3) the number of queries is large, the large amount of noise

↓

schemes:
1) coupled sensitivity ✓
2) save the privacy budgets in the data releasing algorithm
3) coupled iteration mechanism.

5.1
(1) Related work
1) coupled Differential Privacy
2) coupled Data analysis.

research issues:
How to identify coupled records in a dataset?
How to calibrate the sensitivity for coupled records
How to design a coupled data releasing mechanism?

(2) Problem statement:

(3) coupled differential privacy Problem.

1) notion: coupled dataset

△ coupled degree: represents the impact of a record on another record.

coupled degree matrix

2) research issues and challenges

How to identify coupled records in a dataset?
How to calibrate sensitivity for coupled records?
How to re-design the differential privacy mechanism?

5.2 coupled dataset analysis

(1) coupled analysis: to generate coupled degree matrix Δ

1) Attribute analysis: to discover the relationship among records

2) time interval analysis: pre-defines the size of a time interval to identify the correlation.

3) Pearson correlation analysis:

~~Other~~ no proper attribute or time info.

(2) coupled sensitivity

• Definition 1: Record Sensitivity

measures the effect on all records in D when deleting a record r_i

record sensitivity: $CS_i = \sum_{j=0}^{n} |\delta_{ij}| \left(\| Q(D^j) - Q(D^{-j}) \|_1 \right)$, $\delta_{ij} \in \Delta$

δ_{ij} estimates the coupled degree between r_i and r_j

• Definition 2: coupled sensitivity

$$CS_q = \max_{i \in q}(CS_i).$$

q is a record set of all records responding to query Q.

• noise answer:

$$\hat{Q}(D) = Q(D) + Laplace\left(\frac{CS_q}{\varepsilon}\right)$$

(3) coupled Iteration mechanism (CIM)

an iterative-based mechanism will be adopted to limit the noise in the query answer.

(1) overview of algorithm.

true answer: $a_t = Q_t(x)$ ← histogram

noise answer: $\hat{a}_t = Q_t(x) + Laplace\left(\frac{CS_q}{\varepsilon}\right)$

$\hat{d}_t = Q_t(x_{t-1}) - \hat{a}_t$

↓

denotes the difference between the true answer given by x_{t-1} and the noisy answer from x_t.

features:
a. Take relationship between records into consideration
b. decrease the total amount of noise.
c. more queries can be answered than the traditional mechanism with the same privacy budget

(2) coupled update function

Definition 1: coupled update Function

$$x_t(b_i) = x_{t-1}(b_i) \cdot e^{-y \cdot \delta_{q_t} \cdot y_t(x_{t-1})}$$

where $y_t(x_{t-1}) = Q_t(x_{t-1})$ if $\cdot \hat{d} > 0$

} $y_t(x_{t-1}) = 1 - Q_t(x_{t-1})$, otherwise

(3) Parameters disscussion

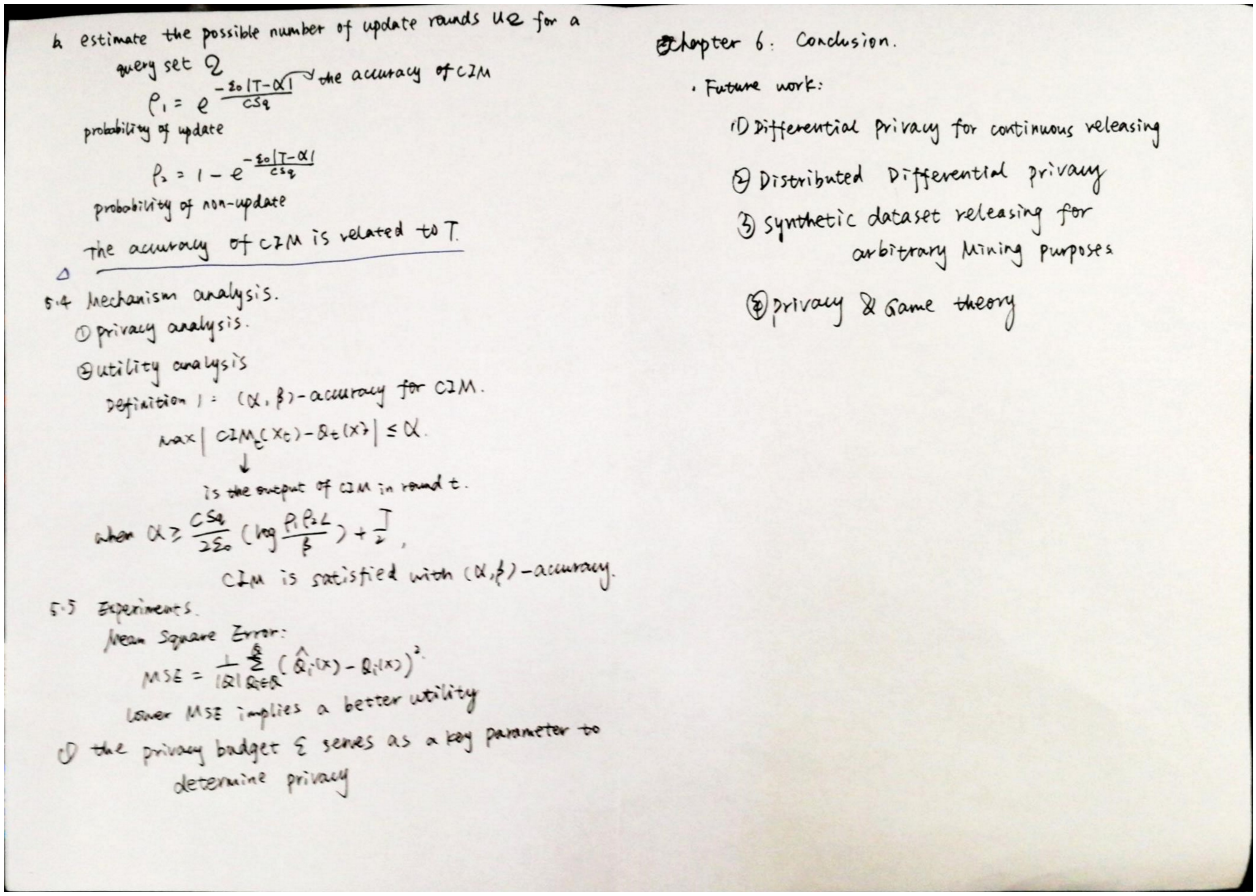
a. measure the maximal number of update rounds U_{max}

$$U_{max} = \frac{\log N}{y^2 \cdot \delta_0^2}$$

y: parameter.

δ_0: couple parameter threshold.

N: the length of a histogram x



estimate the possible number of update rounds uQ for a query set Q

$$\rho_1 = e^{\frac{-\varepsilon_0|T-\alpha|}{cs_q}}$$ ← the accuracy of c2M

probability of update

$$\rho_2 = 1 - e^{\frac{-\varepsilon_0|T-\alpha|}{cs_q}}$$

probability of non-update

the accuracy of c2M is related to T.

△

5.4 Mechanism analysis.

① privacy analysis.

② utility analysis

definition 1: (α, β)-accuracy for c2M.

$$\max\left| c2M_{\varepsilon}(X_t) - q_t(x) \right| \le \alpha.$$

↓ is the output of c2M in round t.

when $\alpha \ge \frac{cs_q}{2\varepsilon_0}\left(\log\frac{\rho_1\rho_2 L}{\beta}\right) + \frac{T}{2}$,

c2M is satisfied with (α,β)-accuracy.

5.5 Experiments.

Mean Square Error:

$$MSE = \frac{1}{|Q|}\sum_{q\in Q}\left(\hat{q}_i(x) - q_i(x)\right)^2.$$

Lower MSE implies a better utility

① the privacy budget ε serves as a key parameter to determine privacy

Chapter 6. Conclusion.

· Future work:

① Differential privacy for continuous releasing

② Distributed Differential privacy

③ synthetic dataset releasing for arbitrary Mining purposes

④ privacy & Game theory

Summary of differential privacy

Summary & Logic

Problems

① high sensitivity queries → large amount of noise → decrease the utility of dataset

② sparsity of the dataset → induce redundant noise.

③ coupled records will release more info (privacy data).

schemes

① application-aware sensitivity

② using randomized mechanism & shrink the randomized domain.

③ remeasure the sensitivity to capture the relation among records
- a. coupled sensitivity
- b. iteration based coupled releasing mechanism

Differential privacy scenarios:

- Data collection
  - --|-- trust boundary
- Data analysis
  - data release
    - interactive setting
    - non-interactive setting
  - data mining
    - interface Based data mining
    - Fully access data mining

Theory: notion; (ε, τ)-Differential privacy
- privacy budget
- violation probability

Sensitivity
- global sensitivity
- local sensitivity

Mechanism
- Laplace – numeric
- Exponential – non-numeric

Differentially Private data release:

- interactive setting
  - mechanism:
    - naive Laplace
    - Query separation
    - Iteration
  - Histogram release
    - partition
    - kd-tree based partition
    - optimal partition method
    - maintained the consistency
- non-interactive setting
  - batch queries release
  - group-base release
  - contigency table release
  - Synthetic dataset release

Differential private Data Mining:

- Data mining with Interface
  - Sub-Linear Queries (SuLQ)
  - Privacy Integrated queries platform (PINQ)
- Data Mining with full access
  - ID3.
  - Logistic regression.
  - frequent Itemset Mining (DiffFIM).

☆ Applications of differential privacy

- recommender system
- tagging recommender system
- coupled differential privacy

1. Differential private neighborhood-based collaborative Filtering (PNCF).

Similarity: { Pearson correlation coefficient (PCC)
cosine-based Similarity (COS)

rating selection { item-based
user-based.

△ PNCF { Private neighbor Selection
Perturbation: exponential mechanism.

---

2. Differential Private Tagging systems

PriTop { Private Topic Model generation { LDA Model construction
private Model Generation
Topic-based profile Generation.
Topic weight Perturbation
Private tag selection.

---

3. coupled differential privacy for non-IID data sets

coupled differential privacy:
☆ coupled Iteration mechanism (CIM)