【ICDE2015勉強会】

Session 22:

Data Privacy and Security 2

担当： 曹 洋 D2（京都大学　吉川・馬研究室）

Outline

1. A Hybrid Private Record Linkage Scheme: Separating Differentially Private Synopses From Matching Records

- Jianneng Cao (Institute for Infocomm Research), Fang-Yu Rao (Purdue University), Elisa Bertino (Purdue University), Murat Kantarcioglu (University of Texas at Dallas)

2. Conservative or Liberal? Personalized Differential Privacy

- Zach Jorgensen (North Carolina State University), Ting Yu (North Carolina State University), Graham Cormode (University of Warwick)

3. Differentially Private Frequent Sequence Mining via Sampling-based Candidate Pruning

- Shengzhi Xu (Beijing University of Posts and Telecommunications), Sen Su (Beijing University of Posts and Telecommunications), Xiang Cheng (Beijing University of Posts and Telecommunications), Zhengyi Li (Beijing University of Posts and Telecommunications), Li Xiong (Emory University)

Differential Privacy

- ε-差分プライバシ　(ε-Differential Privacy, ε-DP)
  - Privacy = “Right to be let alone”
    - database D : each record is an individual’s data
    - user : “Please to analyze D except my data”

D

User	Disease
u1 	心臓病
u2 	HIV
u3 	HIV

Differential Privacy

- ε-差分プライバシ　(ε-Differential Privacy, ε-DP)
  - Privacy = “Right to be let alone”
    - database D : each record is an individual’s data
    - user : “Please to analyze D except my data”
  - DP is a formal definition of this idea
    - For Database **D**, Query **Q;** given **ε**>0;
    - **D’**(**except one individual’s data**) is denoted as “Neighboring Database” of D.

D

User	Disease
u1	心臓病
u2	HIV
u3	HIV

Neighboring Database D’

Database D

Differential Privacy

- ε-差分プライバシ　(ε-Differential Privacy, ε-DP)
  - Privacy = “Right to be let alone”
    - database D : each record is an individual’s data
    - user : “Please to analyze D except my data”
  - DP is a formal definition of this idea
    - For Database **D**, Query **Q;** given **ε**>0;
    - **D’**(except one individual’s data) is denoted as “Neighboring Database” of D.
    - If **Mechanism M** guarantees result ***r*** of Q(D) and Q(D’) are “similar”:

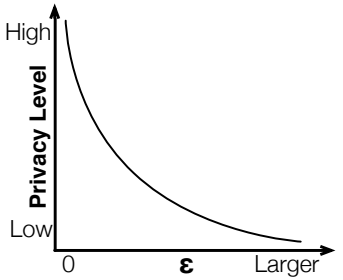
D

User	Disease
u1	心臓病
u2	HIV
u3	HIV

$$\frac{\Pr(M(Q(D))=r)}{\Pr(M(Q(D'))=r)} \le e^{\varepsilon}$$

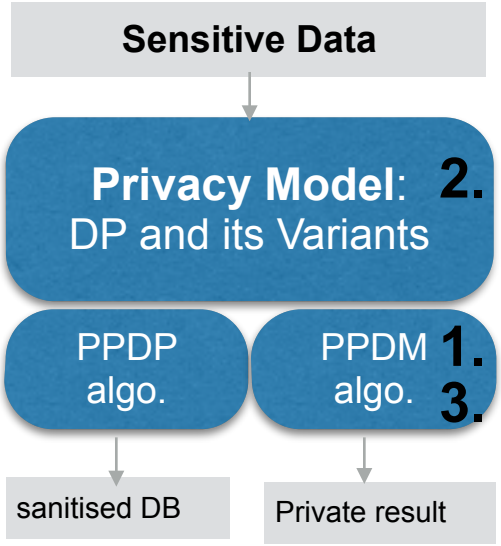
e.g., $\frac{\Pr(M(Q(\text{👥}))=r)}{\Pr(M(Q(\text{👥}))=r)} \le e^{\varepsilon}$

    - then M satisfies ε-DP. ε is a specified privacy level:



Differential Privacy

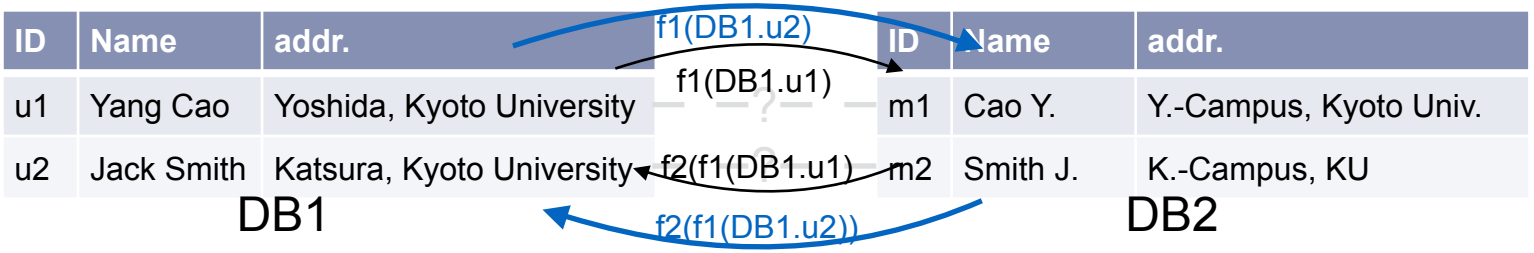
- ε-差分プライバシ (ε-Differential Privacy, ε-DP)
  - DP-based data analysis
    - Privacy Preserving Data Publishing (PPDP)
      - sensitive data → private (“safer”) data
    - Privacy Preserving Data Mining (PPDM)
      - DM algo. ⟹ DP-algo
      - sensitive data →private result



(research scope)

1. A Hybrid Private Record Linkage Scheme:
Separating Differentially Private Synopses From Matching Records
2. Conservative or Liberal? Personalized Differential Privacy
3. Differentially Private Frequent Sequence Mining via Sampling-based Candidate Pruning

- Background:
- Privacy Preserving Record Linkage (PPRL)
  - multiple data owners want to know their common records (RL)
  - but the **data is too sensitive to be exchanged** (PPRL)
  - (resource) a tutorial of PPRL [*]

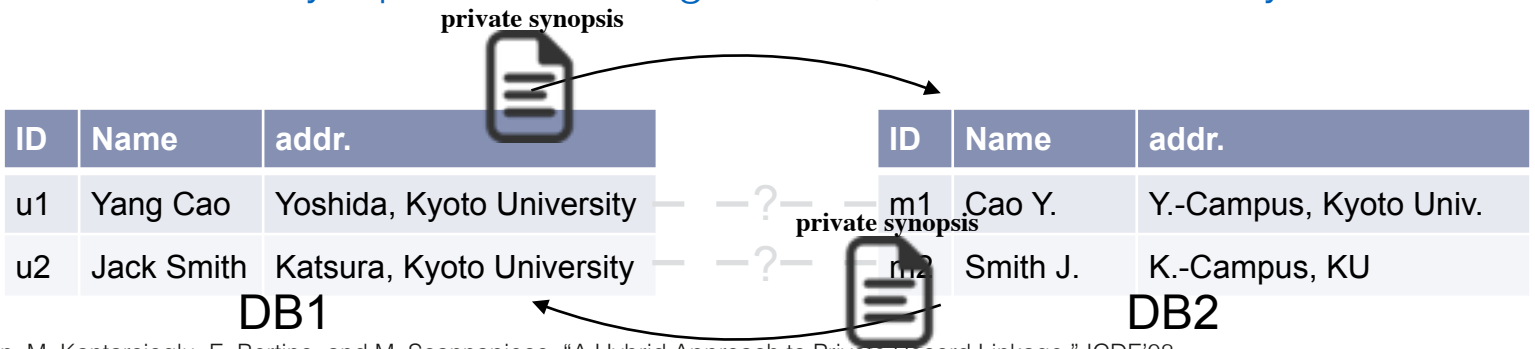


- Secure Multi-party Computation (SMC)-based PPRL
  - SMC:make sure each party **only** learn about final result
  - In PPRL, SMC approach is used to **calculate similarities of possible matches** (based on homomorphic encryption)

[*]Peter Christen ,et al., Techniques for Scalable Privacy-preserving Record Linkage, tutorial on CIKM'13,

Motivation

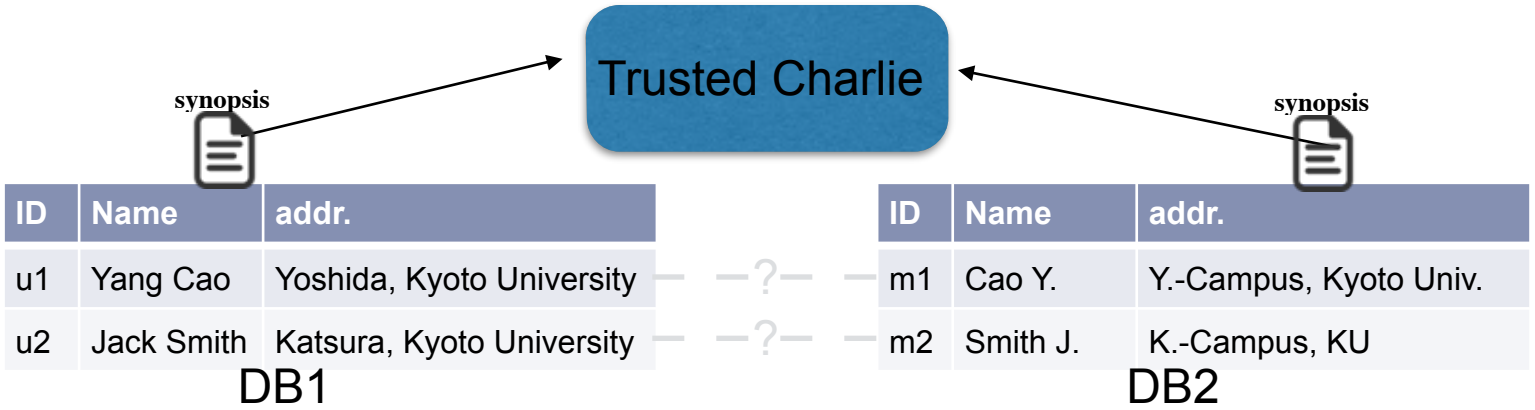
- drawback of SMC-based PPRL
  - **O(m*n)** computational time (m,n are size of two owners' DB)
  - ⟹Hybrid approach [1,2](SMC+private synopsis) to make it more scalable:
    - 1) release **private synopsis** ([1] by k-anonymity [2] by DP) to each other
      - **synopses**: Spatial indexing techniques (BSP-Tree, KD-Tree, or R-Tree) are used to form sub-sets (hyper-rectangles)
    - 2) compute similar records by private synopsis
  - But [2] is **not totally differential private**! [this paper proved] (省略)
    - when DP-synopsis + matching records, DP do not hold anymore



[1] A. Inan, M. Kantarcioglu, E. Bertino, and M. Scannapieco, "A Hybrid Approach to Private Record Linkage," ICDE'08
[2] A. Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino, "Private record matching using differential privacy," in EDBT'10

1. A Hybrid Private Record Linkage Scheme:
Separating Differentially Private Synopses From Matching Records

- Proposed solution
  - Separating Differentially Private Synopses From Matching Records
  - two parties:Alice,Bob (previous study)⇒three parties:+TrustedCharlie (proposed)
    - Charlie prunes the record matching based on the received synopses.
  - Building blocks
    - Secure Multi-party Computation (based on *Paillier Cryptosystem*)
    - Differential Private synopsis
- Contribution
  - approximate PPRL framework
  - prove the safety of this framework (satisfy differential privacy & SMC)



2. Conservative or Liberal? Personalized Differential Privacy

Recall: ε_i ~ privacy level
smaller ε ~ higher privacy
larger ε ~ lower privacy

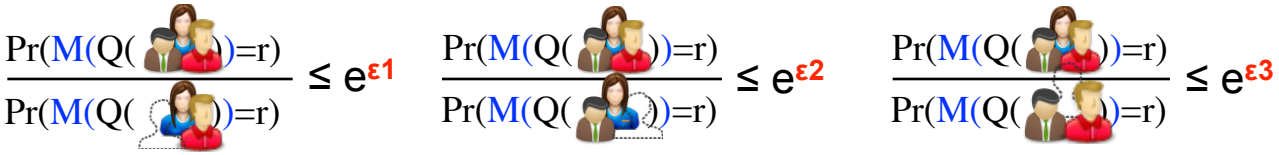
- **Motivation**
  - In differential privacy, **all users** in dataset receive **the same privacy guarantee**
    - if $\frac{Pr(M(Q(\text{all}))=r)}{Pr(M(Q(\text{all}-1))=r)} \le e^{\varepsilon 1}$ $\frac{Pr(M(Q(\text{all}))=r)}{Pr(M(Q(\text{all}-2))=r)} \le e^{\varepsilon 2}$ $\frac{Pr(M(Q(\text{all}))=r)}{Pr(M(Q(\text{all}-3))=r)} \le e^{\varepsilon 3}$

      DP only consider "worst case"

      then M need to satisfies ε_{min}-DP (ε_{min}=**Min**{$\varepsilon 1, \varepsilon 2, \varepsilon 3$}) :

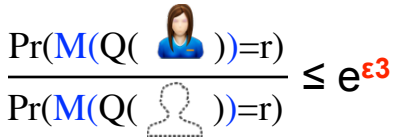
      $\frac{Pr(M(Q(\text{all}))=r)}{Pr(M(Q(\text{all}-1))=r)} \le e^{\varepsilon\_min}$ $\frac{Pr(M(Q(\text{all}))=r)}{Pr(M(Q(\text{all}-2))=r)} \le e^{\varepsilon\_min}$ $\frac{Pr(M(Q(\text{all}))=r)}{Pr(M(Q(\text{all}-3))=r)} \le e^{\varepsilon\_min}$
  - In reality, **different individuals** often have **different privacy preferences**
    - some users are under-protected, while others are over-protected.
- How to make "**Personlized DP**" possible ? (exactly satisfy ε_i)

- Proposed Mechanisms for PDP



  - *Baseline*
    - achieve **Min**{ε1,ε2,ε3}-DP, because of DP ⇒PDP
  - *Threshold*
    - give a threshold $\varepsilon_{min} \le \mathbf{t} \le \varepsilon_{max}$, delete all data of u whose $\varepsilon_u < \mathbf{t}$
    - then achieve **t**-DP by traditional DP mechanisms

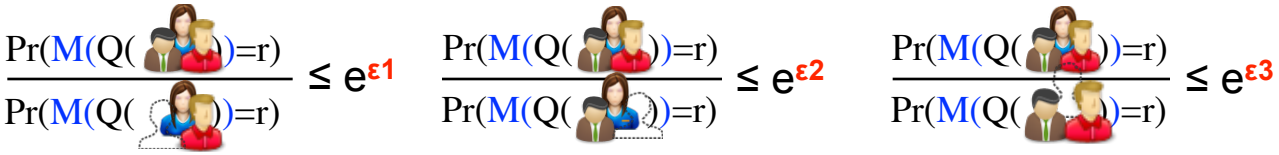
例
in the above example:
given ε1=1,ε2=2,ε3=3, t=2.5,
By *threshold* **mechanism**, first deleting data of u1,u2
then only need to achieve :



D

User	Disease
~~u1~~	~~心臓病~~
~~u2~~	~~HIV~~
u3	HIV

- Proposed Mechanisms for PDP



  - *Baseline*
    - achieve **Min**{ε1,ε2,ε3}-DP, because of DP ⇒PDP
  - *Threshold*
    - give a threshold $\varepsilon_{min} \le \mathbf{t} \le \varepsilon_{max}$, delete all data of u whose $\varepsilon_u < \mathbf{t}$
    - then achieve **t**-DP by traditional DP mechanisms
  - *Sample*
    - give a threshold $\varepsilon_{min} \le \mathbf{t} \le \varepsilon_{max}$, "carefully" sample a subset of D
    - then achieve **t**-DP by traditional DP mechanisms
  - *Exponential-like*
    - essentially, it is a **modified Exponential Mechanism**

2. Conservative or Liberal? Personalized Differential Privacy

- **Contributions:**
- Formally define a **new privacy model** based on DP: **Personalized Differential Privacy (PDP).**
  - each u_i has a privacy preference ε_i
  - PDP is a **generalization** of DP: algo A satisfy DP ⇒ A satisfy PDP.
- Proposed serval **Mechanisms** to achieve PDP
  - any existing DP-algorithm can be transformed to PDP-algorithm by these mechanisms
- **Experiments** on real/artificial dataset to transform serval simple DP-algorithms to PDP-algorithms
  - PDP-Median (synthetic data) *Sample.M* is best
  - PDP-Count (synthetic data) *Exponential-like.M* is best
  - PDP-Linear regression (2012 U.S. census data) *Sample.M* is best

Motivation

- Frequent Sequence Mining (FSM) - a fundamental DM algo.

ID	sequence
seq1	abc
seq2	abcd
seq3	abcde
seq4	abe

A sequence DB

Threshold=3

support

	frequent sequences
3-seq	abc(3)
2-seq	ab(4),ac(3),bc(3)
1-seq	a(4),b(4),c(3)

C3={abc,bac,…}

C2={ab,ac,bc,ba..}

FSM result

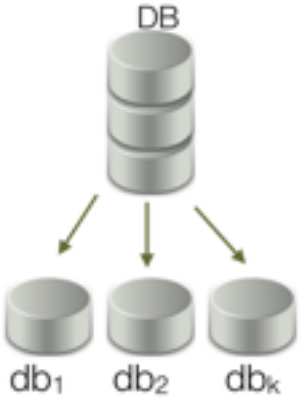
- **When sequences is sensitive : DP version of FSM** (DP-FSM)

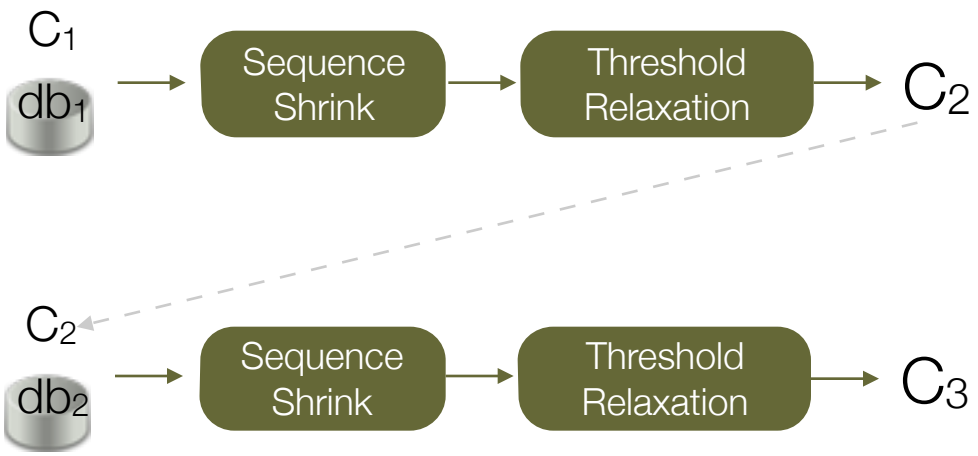
- How to design **DP-FSM**?

$$\frac{\Pr(\text{DP-FSM}(\text{D}_1)=r)}{\Pr(\text{DP-FSM}(\text{D}_2)=r)} \le e^{\varepsilon}$$

  - problem: e.g., removing *seq3* will significantly effect 1-seq,2-seq…
  - idea: reduce candidates of i-seq→(i+1)-seq , author use **heuristic methods** to detect which i-seq is potentially impossible (i+1)-seq.

- Proposed Solution





....

  - **Sequence Shrink**
    - Irrelevant item detection
    - Consecutive Pattern Compression
    - Sequence Reconstruction
  - **Threshold Relaxation**
    - decrease the probability of misestimating in db_k

- **Related work:**
- **Frequent Consecutive Sequence Mining[*]**
  - difference:
    - 1) [*] consider consecutive
    - 2) [*] is interactive publish, this paper is non-interactive publish

publish APIs

publish sanitised DB

- Line of work in DP-based Frequent Pattern Mining
  - DP-Frequent Graph Mining (FGM) [3]
  - DP-Frequent Item Mining (FIM) [4,5]
  - DP-Frequent Sequence Mining (FSM) [6]

[*]L. Bonomi and L. Xiong, "A Two-phase Algorithm for Mining Sequential Patterns with Differential Privacy," CIKM,2013.
[3] E. Shen and T. Yu, "Mining Frequent Graph Patterns with Differential Privacy," KDD 2013.
[4] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta, "Discovering Frequent Patterns in Sensitive Data," KDD, 2010.
[5] R. Chen, et al., "Publishing set-valued data via differential privacy," VLDB, 2011.
[6] R. Chen, et al., "Differentially Private Sequential Data Publication via Variable-Length N-Grams," ACM CCS, 2012.