k-Support Anonymity Based on Pseudo Taxonomy for Outsourcing of Frequent Itemset Mining

Chih-Hua Tai [+,*], Philip S. Yu [*] and Ming-Syan Chen [+,&]

[+]Dept. of EE, National Taiwan University, Taiwan

[*]Dept. of CS, University of Illinois at Chicago, USA

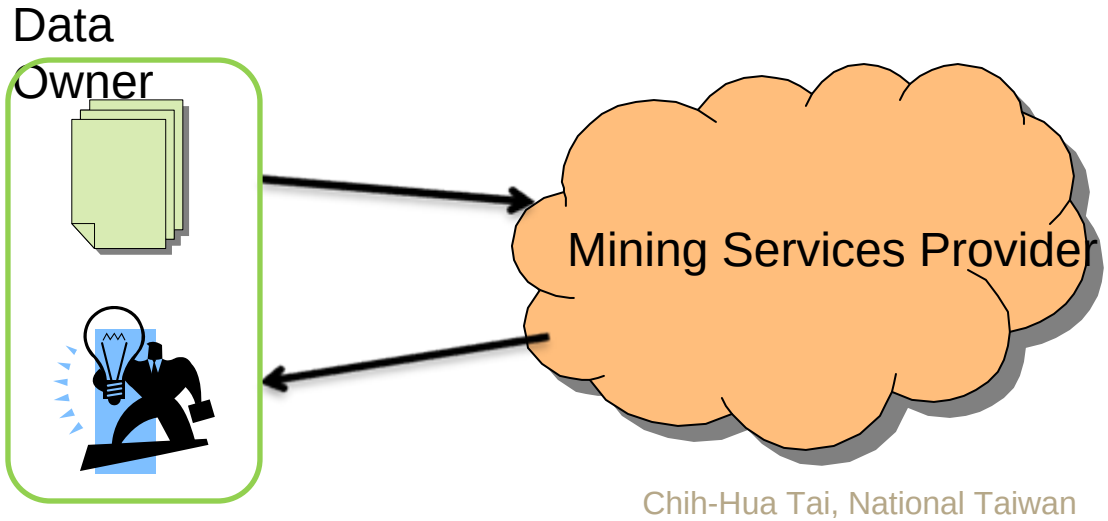
[&]Research Center for Information Technology Innovation, Academia Sinica, Taiwan

Outlines

- The needs, the problems and the challenges of outsourcing of frequent itemset mining (FIM)
- Related works
- K-support anonymization
- Anonymization algorithm (Encryption/Decryption method)
- Performance studies
- Conclusions

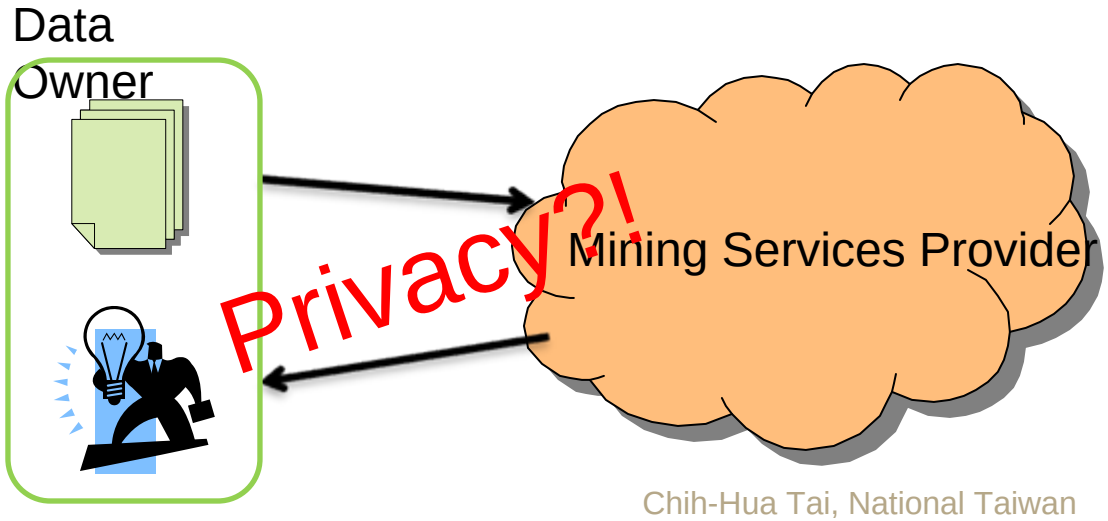
The needs of outsourcing FIM

- Frequent itemset mining (FIM) has shown its great promise in various fields.
  - Business, Medical treatment, Networks, and Bioinformatics
- For those who lack of expertise in FIM and/or computing resources, they have the need of outsourcing the mining tasks to a professional third party.



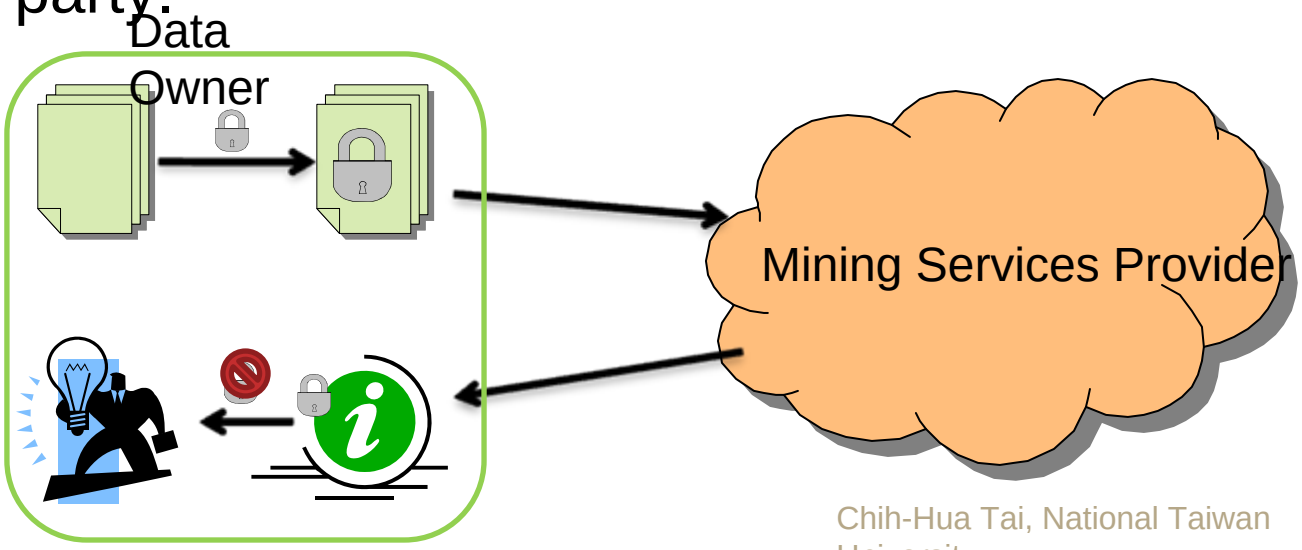
The problems of outsourcing FIM

- Frequent itemset mining (FIM) has shown its great promise in various fields.
  - Business, Medical treatment, Networks, and Bioinformatics
- For those who lack of expertise in FIM and/or computing resources, they have the need of outsourcing the mining tasks to a professional third party.



The problems of outsourcing FIM

- Frequent itemset mining (FIM) has shown its great promise in various fields.
  - Business, Medical treatment, Networks, and Bioinformatics
- For those who lack of expertise in FIM and/or computing resources, they have the need of outsourcing the mining tasks to a professional third party.



The challenges of outsourcing FIM

- The mining results should remain correct and complete.
- Privacy in both raw data and mining results should be effectively protected.
- The overhead of encryption and decryption should be reasonable.

The challenges of outsourcing FIM

Trans. ID	Items
1	wine
2	cigar, wine
3	cigar, tea
4	beer, cigar, wine
5	beer, tea, wine

Encrypt →

Trans. ID	Items
1	a
2	a, c
3	c, d
4	a, b, c
5	a, b, d

The challenges of outsourcing FIM

- Top frequency attack
  - Wine is the most frequent item → ‘a’ is ‘wine’
- Approximate support attack
  - The support of cigar is about 55%~60% →‘c’ is ‘cigar’

Trans. ID	Items
1	wine
2	cigar, wine
3	cigar, tea
4	beer, cigar, wine
5	beer, tea, wine

Encrypt →

Trans. ID	Items
1	a
2	a, c
3	c, d
4	a, b, c
5	a, b, d

The challenges of outsourcing FIM

The support information about the frequent itemsets can be utilized to effectively reveal the raw data as well as the sensitive information from the anonymized transactions.

T. Mielik¨ainen. Privacy problems with anonymized transaction databases. In *Proc. of Discovery Science,* 2004.

Trans. ID	Items
1	wine
2	cigar, wine
3	cigar, tea
4	beer, cigar, wine
5	beer, tea, wine

Encrypt →

Trans. ID	Items
1	a
2	a, c
3	c, d
4	a, b, c
5	a, b, d

Related Works

- Encrypt each real items by a one-many mapping function.

  Wong, W. K., Cheung, D. W., Hung, E., Kao, B., Mamoulis, N.: Security in Outsourcing of Association Rule Mining. In: Proc. of VLDB, 2007.

- However, it does not try to anonymize the support information.
- Recently it is cracked.

  Molloy, I., Li, N., Li, T.: On the (In)Security and (Im)Practicality of Outsourcing Precise Association Rule Mining. In: Proc. of ICDM, 2009.

K-support Anonymity

- For every sensitive item, there are at least k-1 other items of the same support.

- Because of k-support anonymity, an experience attacker cannot succeed in re-identifying sensitive items even with the precise support information.

Problem Formulation

- Given a transactional database T, encrypt T into T' such that
  - There exist a decryption function *D* such that MiningResult(T Δ)= *D*(MiningResult(T', Δ)), for any minimal support Δ.
  - T' is k-support anonymous.

A naïve solution

- For each set of real items of the same support, add enough fake items randomly into transactions to make the fake items as frequent as real ones.

Trans. ID	Items
1	wine
2	cigar, wine
3	cigar, tea
4	beer, cigar, wine
5	beer, tea, wine

Items
a, e, g, h, i
a, c, e, f, h, i
c, d, e, f, g
a, b, c, f, h
a, b, d, e, f, g

For k = 3,
16 additional items are required.

4 x 2 = 8 (e, f) for wine
3 x 2 = 6 (g, h) for cigar
2 x 1 = 2 (i) for beer and tea

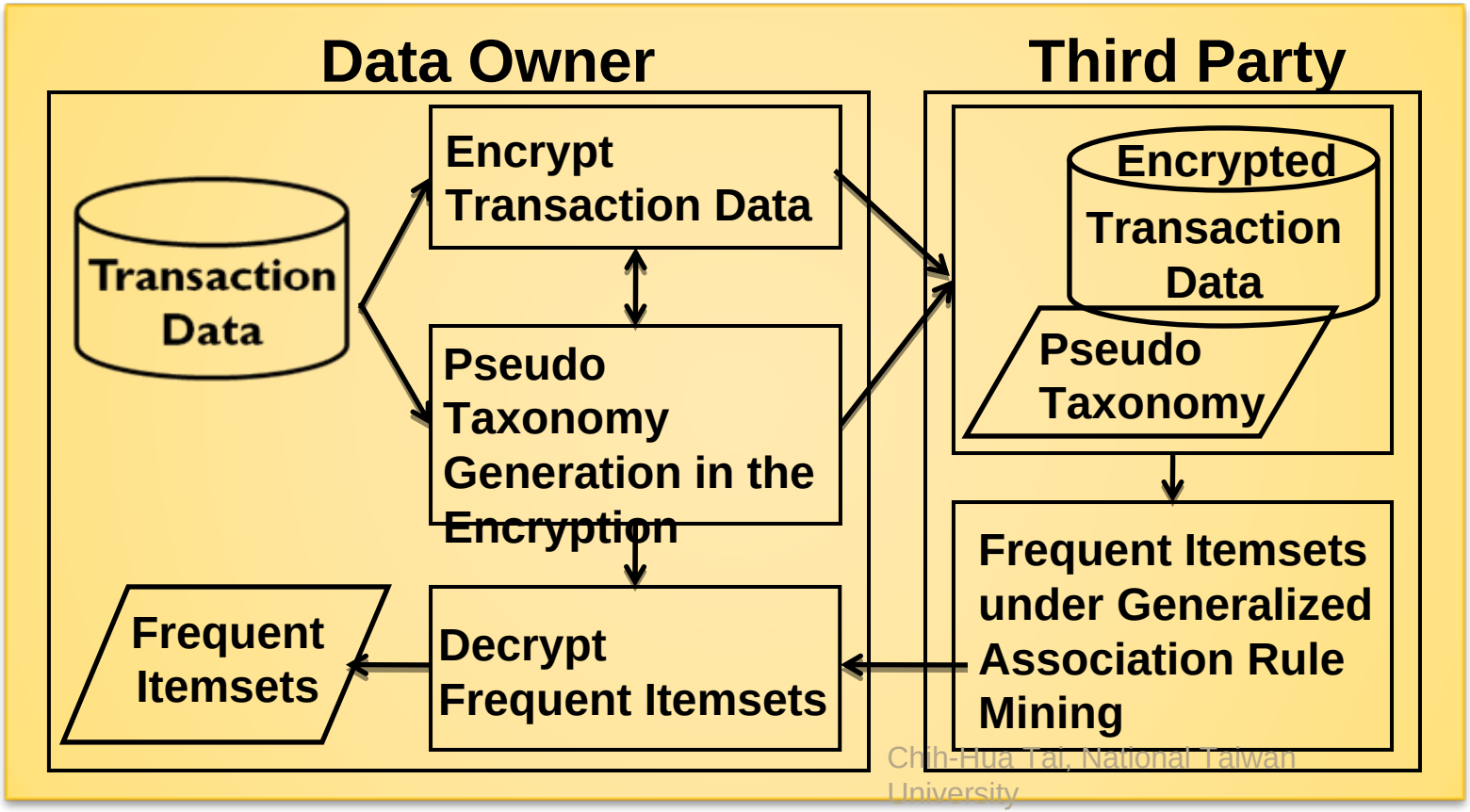
- However, there could be too large storage overhead when k is large.

Anonymization: Overview

- For storage efficiency, we suggest to convert FIM to Generalized FIM, which discovers all frequent itemsets across levels of a given taxonomy.

Anonymization: Overview

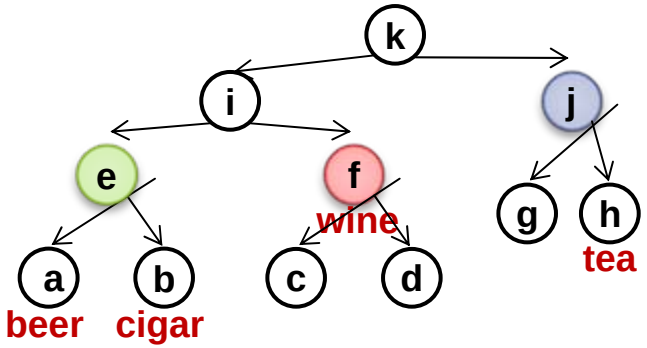
- For storage efficiency, we suggest to convert FIM to Generalized FIM, which discovers all frequent itemsets across levels of a given taxonomy.



Anonymization: Overview

- Under generalized association rule mining, items can be at multiple levels of a taxonomy and only the items at leaf level need to appear in the database.

Trans. ID	Items
1	wine
2	cigar, wine
3	cigar, tea
4	beer, cigar, wine
5	beer, tea, wine



Encrypt with k=3

Trans. ID	Items
1	c, d, g
2	b, d, g
3	b, h
4	a, b, c
5	a, c, d, h

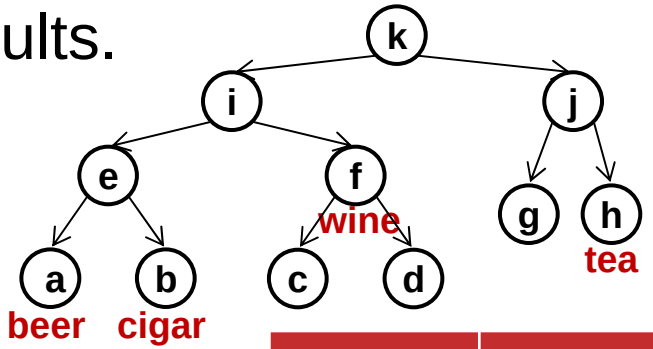
2
1
1

wine ∈ {e, f, j}
cigar ∈ {b, c, d}
beer and tea ∈ {a, g, h}

4 additional items required

Anonymization: Overview

- The real frequent itemsets can be obtained by filtering out patterns containing any fake item in 1 scan of the returned results.



Trans. ID	Items
1	wine
2	cigar, wine
3	cigar, tea
4	beer, cigar, wine
5	beer, tea, wine

Trans. ID	Items
1	c, d, g
2	b, d, g
3	b, h
4	a, b, c
5	a, c, d, h

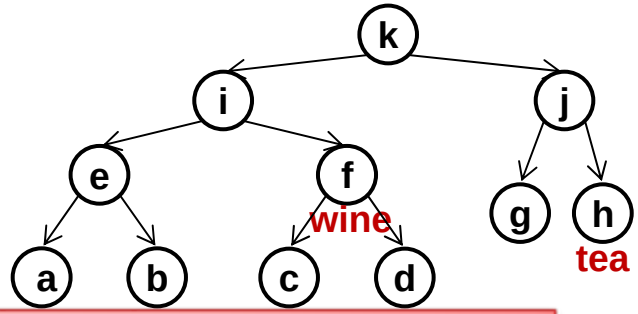
min_sup = 2

Result={{beer}, {cigar}, {wine}, {tea},
{beer, wine}, {cigar, wine}}

Result={a, b, c, d, e, f, g, h, i, j, k,
ac, af, bf, ce, …}

Anonymization: Overview

Trans. ID	Items
1	wine
2	cigar, wine
3	
4	
5	beer, tea, wine



The problem is how to build the taxonomy and encrypt T for k-support anonymity.

2	b, d, g
3	b, h
4	a, b, c
5	a, c, d, h

Anonymization: Overview

- 1: Generalization of the Mining Task
  - To generate a pseudo taxonomy that can

    (a) conserve the correct and complete mining results,

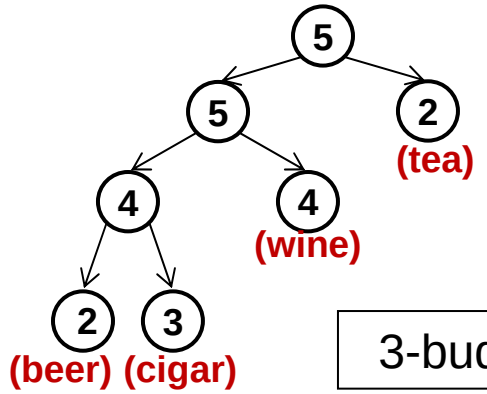
    (b) facilitate k-support anonymization.
- 2: Anonymization with Taxonomy Tree
  - To encrypt T for k-support anonymity with the help of the constructed taxonomy tree.

1: Generalization of the Mining Task

- Build a k-bud tree of T
  - All real items at the leaf level
  - The number of nodes in three categories is equal to or greater than k

    Let x_M denote the most frequent real item in T
    - $A_> = \{ v \mid sup(v) > sup(x_M)$ and v is leaf$\}$,
    - $A_= = \{ v \mid sup(v) = sup(x_M)\}$, and
    - $A_< = \{ v \mid sup(v) < sup(x_M) < sup(u)$, where u is the parent node of v $\}$.

Trans. ID	Items
1	wine
2	cigar, wine
3	cigar, tea
4	beer, cigar, wine
5	beer, tea, wine



3-bud tree

2: Anonymization with Taxonomy Tree

3-support anonymity

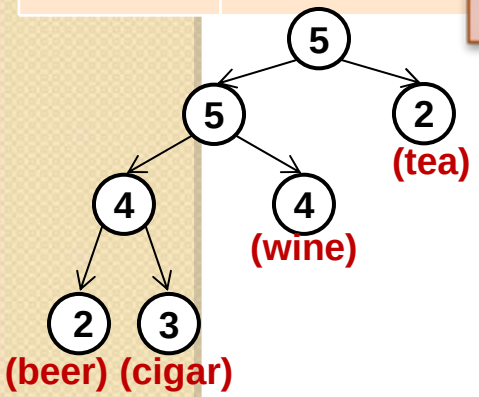
Trans. ID	Items
1	wine
2	cigar, wine
3	cigar, tea
4	beer, cigar, wine
5	beer, tea, wine

Items
wine, p1
cigar, wine, p1
cigar, tea
beer, cigar, wine
beer, tea, wine

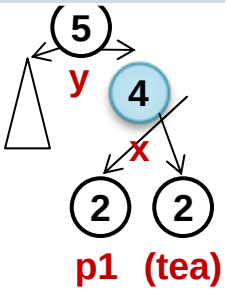
Items
p1, p2
cigar, p1, p3
cigar, tea
beer, cigar, p2
beer, tea, p2

Items
p1, p2, p3
cigar, p1, p3
cigar, tea
beer, cigar, p2
beer, tea, p2, p3

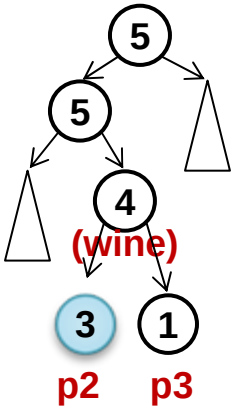
Trans. ID	Items
1	c, d, g
2	b, d, g
3	b, h
4	a, b, c
5	a, c, d, h



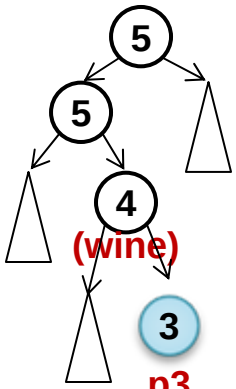
3-bud tree



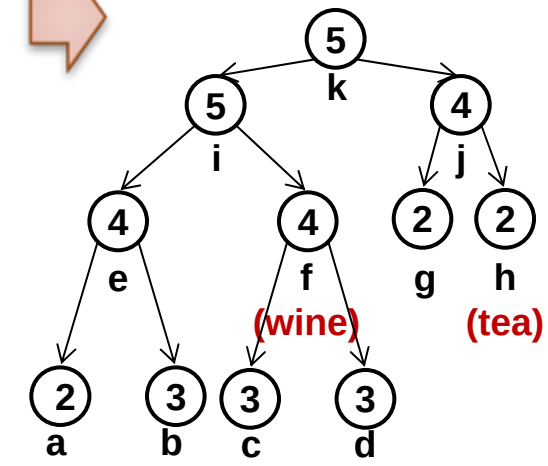
insertion



split



increase

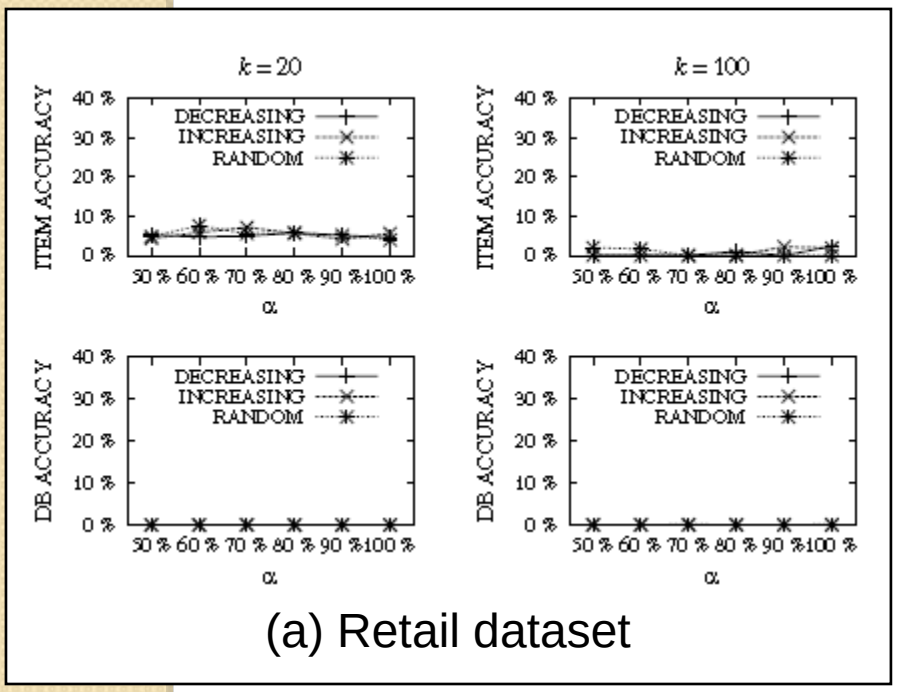


Performance Studies

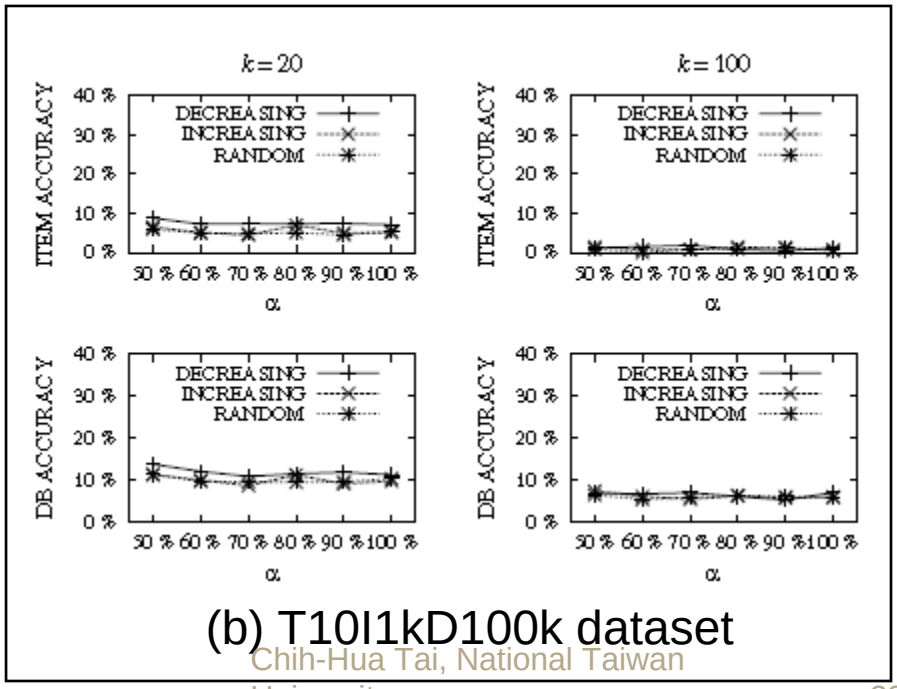
- Data sets
  - Retail dataset
    - 88162 transactions with 2117 different items
  - T10I1kD100k dataset
    - 100k transactions with 1000 different items

Performance Studies

- Protection against precise item support attack
  - Item accuracy
    - The ratio of items being re-identified
  - DB accuracy
    - The avg. ratio of items in a transaction being re-identified



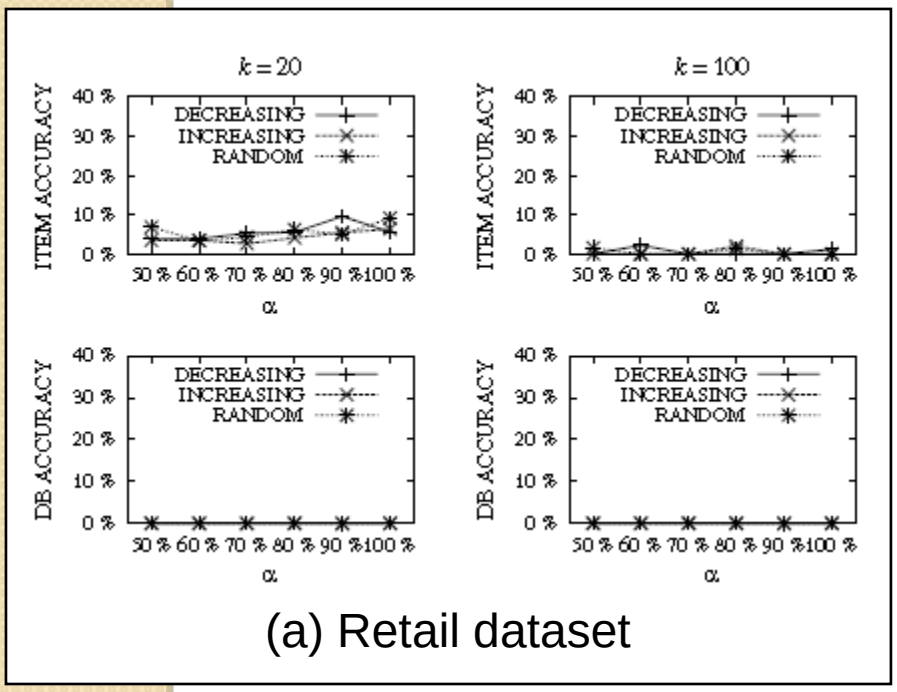
(a) Retail dataset



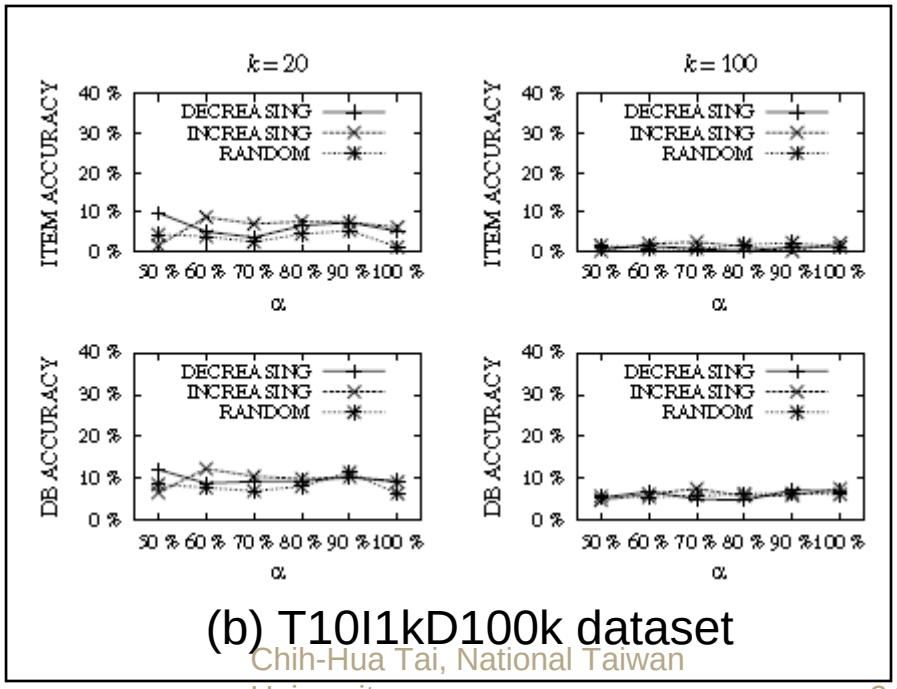
(b) T10I1kD100k dataset

Performance Studies

- Protection against precise itemset support attack
  - Item accuracy
    - The ratio of items being re-identified
  - DB accuracy
    - The avg. ratio of items in a transaction being re-identified



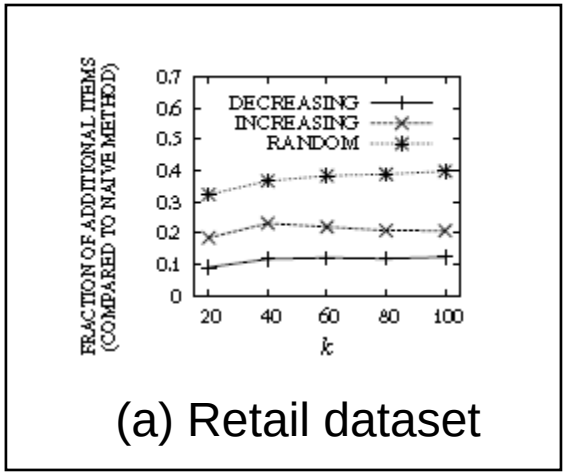
(a) Retail dataset



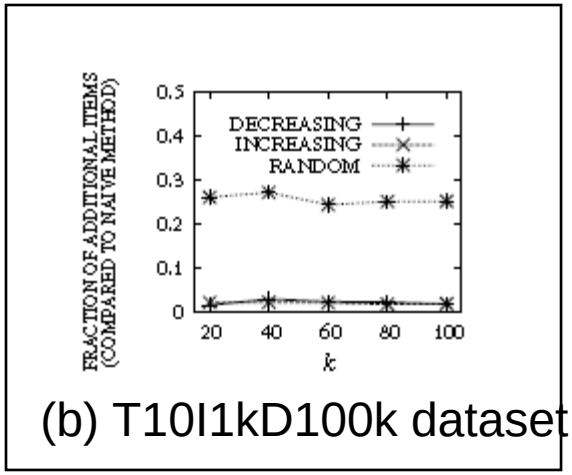
(b) T10I1kD100k dataset

Performance Studies

- Storage overhead



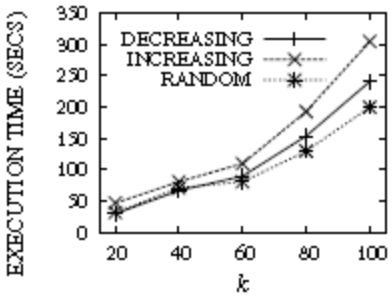
(a) Retail dataset



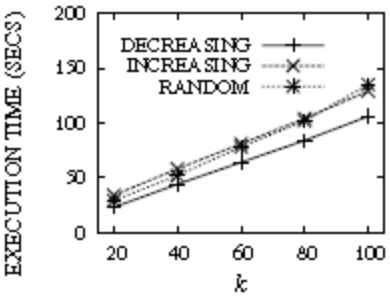
(b) T10I1kD100k dataset

Performance Studies

- Execution efficiency



(a) Retail dataset



(b) T10I1kD100k dataset

Conclusions

- We proposed k-support anonymity to enhance the privacy protection in outsourcing of frequent itemset mining (FIM).
- For storage efficiency, we transformed FIM to Generalized FIM, and proposed a taxonomy-based anonymization algorithm.
- Our method allows the data owner to obtain the real frequent itemsets in1 scan of the returned results.
- Experimental results on both real and synthetic data sets showed that our method can achieve very good privacy protection with moderate storage overhead.

THANK YOU~!

Q & A