Privacy-preserving Frequent Itemset Mining

Raghav Bhaskar[1], Srivatsan Laxman[1], Adam Smith[2] and Abhradeep Thakurta[2]
[1]Microsoft Research, India. [2]Pennsylvania State University.

Research Goal

Frequent Itemset Mining (FIM): Given a database of customer transactions (where each transaction is the number of items a customer buys in one single visit) the objective is to find a list of sets of items which are frequently bought together.

Privacy Notion: To safeguard against leakage of customer information while allowing Frequent Itemset Mining.

Differential Privacy (DMNS'06). *A database access mechanism $\mathcal{A}$ is ϵ-differentially private if for all $x, y \in \mathcal{D}^n$ satisfying $|x \Delta y| = 1$ and for all sets $\mathcal{S}$ of possible outputs*

$$Pr[\mathcal{A}(x) \in \mathcal{S}] \leq e^{\epsilon} Pr[\mathcal{A}(y) \in \mathcal{S}].$$

Solution: We develop a Differentially Private FIM mechanism.

- This provides information theoretic privacy guarantee to individual transactions against information leakage.

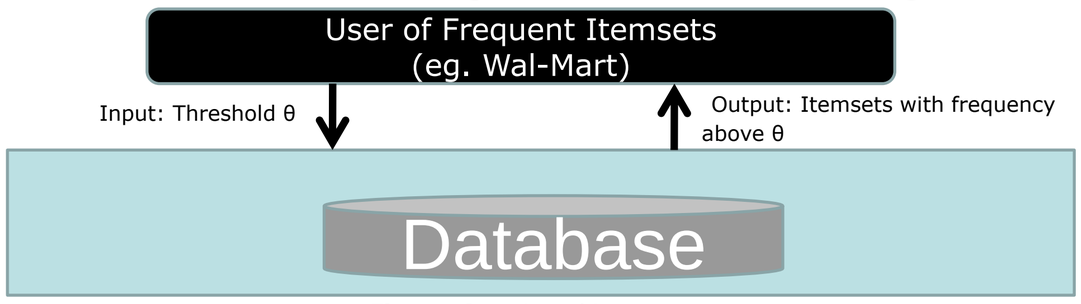
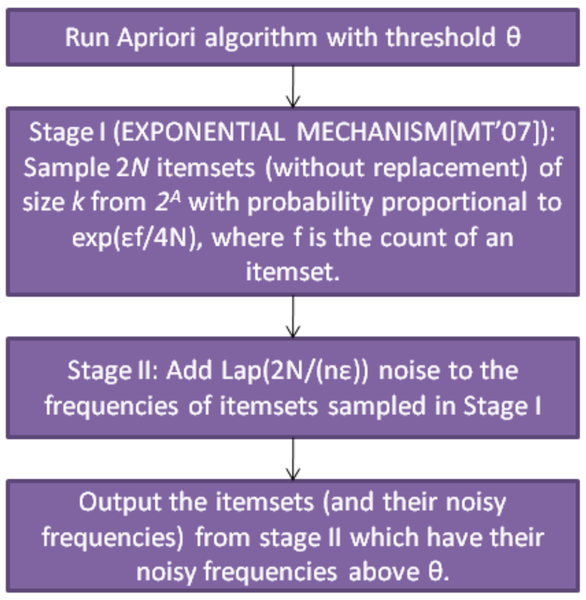


Figure: Generic architecture

Notation

T: Transaction database of size *n.*
A: Item base of *m* items.
Θ: Normalized support threshold.
ε : Privacy parameter.
k: Size of the Frequent Itemsets being mined.
N: Maximum number of itemsets to be output.
freq(.): Array that stores the frequencies of itemsets. Any frequency less than θ is clamped up to θ.

Algorithm



Privacy Guarantee

Theorem. *The algorithm is $2 \cdot \epsilon$-differentially private.*

This means for ε=0.2, contributing information into database increases the probability of any outcome by at most exp(ε)=1.22.

Utility Guarantees

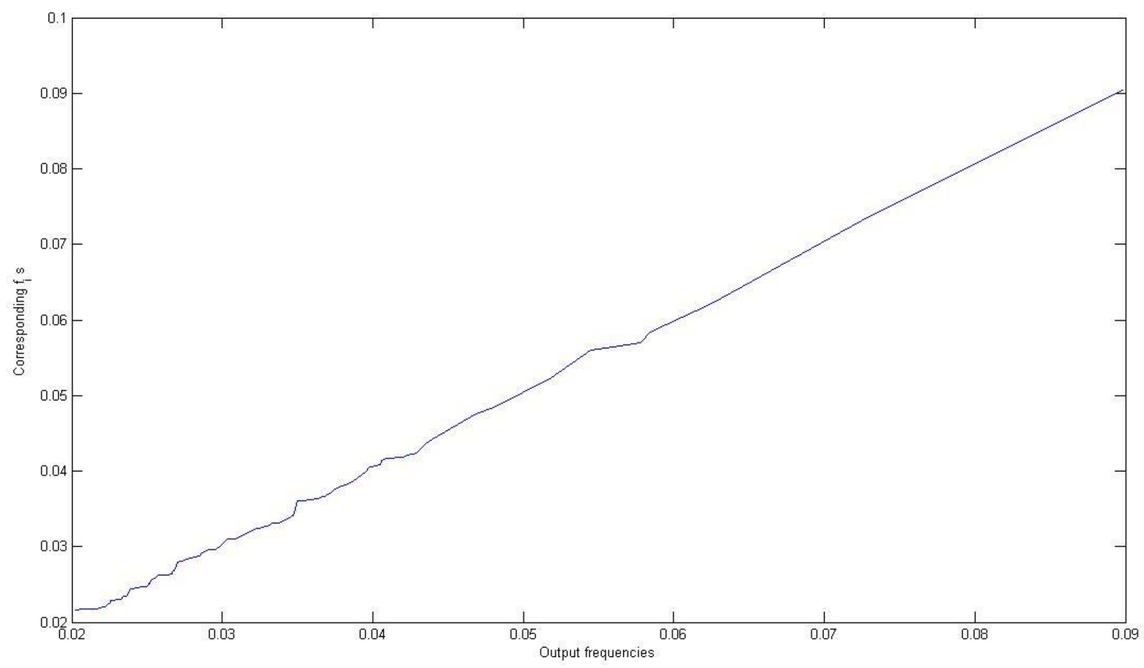
Let f_i denote the frequency of the i-th highest frequency in the array $freq(\cdot)$ and and let $\mathcal{S} = \{\widehat{f}_1, \cdots, \widehat{f}_{2N}\}$ be the frequency of the itemsets sampled in stage I.

Theorem. *For any $0 < \delta < 1$, with probability at least $1 - exp\left(-\frac{N}{8}\left(e^{\frac{\epsilon\delta}{4N}} - 4\right)\right)$, there exist at least N frequencies $\widehat{f} \in \mathcal{S}$ such that, $\widehat{f} \geq f_N - \delta$.*

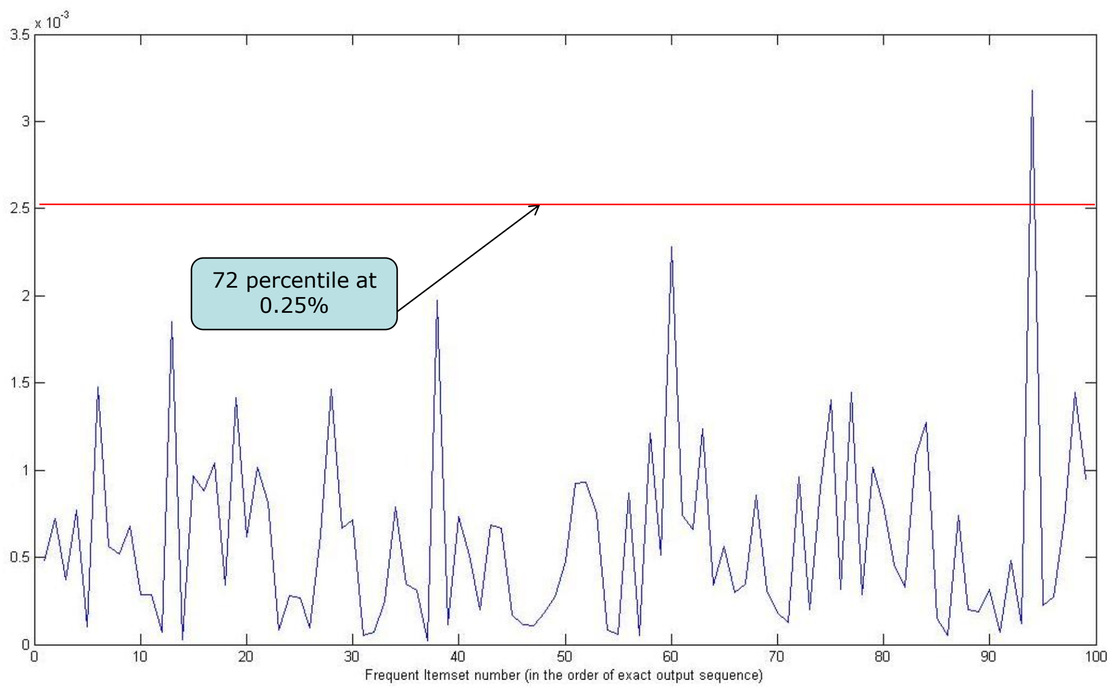
Theorem. *For any $0 < \eta < 1$, with probability at least $1 - N \cdot e^{-\frac{n\epsilon\eta}{2N}}$, all frequencies reported after stage II have their reported frequencies within η margin of their true frequencies.*

Experiments

Parameters: n=970783, m=50, Θ=0.0027, ε=0.2, k=3, N=50 . Dataset generated using IBM synthetic dataset generator.



Plot 1: Actual vs output frequency



Plot 2: Relative error

References

[DMNS'06] C. Dwork, F. McSherry, K. Nissim, A. Smith: Calibrating Noise to Sensitivity in Private Data Analysis. TCC 2006: 265-284

[MT'07] F. McSherry, K. Talwar: Mechanism Design via Differential Privacy. FOCS 2007: 94-103

Project Sponsored by Microsoft Research, India and National Science Foundation