Mining Frequent Patterns Through Microaggregation in Differential Privacy

Z. NI\textsuperscript{1}, Q.M. LI\textsuperscript{1}\textsuperscript{*}, X.Q. LIU\textsuperscript{1}, T.LI\textsuperscript{2}, R.HOU\textsuperscript{3}
\textsuperscript{1}School of Computer Science and Engineering, Nanjing University of Science and Technology No.200 Xiaolingwei, 210094 Nanjing, China
\textsuperscript{2}School of Computer Science, Florida International University 11200 SW 8\textsuperscript{th} Street, Miami, FL 33199, U.S.A.
\textsuperscript{3}Yahoo Center of Latin America, Blanco Encalada 2120, Santiago, Chile
\textsuperscript{*}qianmu@mail.njust.edu.cn

ABSTRACT: Frequent pattern mining has been widely employed to analyze transaction datasets, but the question of how sensitive information contained in a dataset should be protected remains relatively unanswered. The differential privacy model provides a robust privacy guarantee, but the k-anonymity model provides better dataset utility. In this paper, a synergetic approach is proposed to simultaneously protect privacy and enhance data utility when mining top-k frequent patterns. First, microaggregated data is released, which achieves k-anonymity, regardless of the query types the user may be using. Second, top-k frequent patterns are selected based on microaggregated data using the exponential mechanism. Finally, the true support of each top-k frequent pattern is perturbed by adding Laplace noise.

Categories and Subject Descriptors
H.2.8 [Database Applications]: Data mining; K.4.1 [Public Policy Issues]: Privacy

General Terms: Data Mining, Privacy Protection

Keywords: Frequent Pattern, Differential Privacy, K-Anonymity, Microaggregation

Received: 17 December 2014, Revised 24 January 2015, Accepted 29 January 2015

1. Introduction

Frequent pattern mining has been well studied within the data mining context, and numerous frequent pattern mining algorithms have been developed. Among them, the Apriori algorithm [1] and the FP-Growth algorithm [2] are the two most widely used. However, these algorithms are unable to preserve the privacy of the dataset if sensitive information (e.g., patient health records) is contained therein. Bhaskar et al. [3] proposed the Trun-Fre approach based on a Laplace mechanism to achieve differential privacy. However, this does not perform well when candidate sets become overly large, as the length of the frequent pattern increases. To decrease the magnitude of a candidate set, Li et al. [4] developed the PrivBasis approach, under which: (1) The pattern basis is generated based on maximal clique; and (2) top-k patterns are mined to achieve differential privacy. However, the released dataset to be mined often lacks utility. For example, the support of patterns \{a\} and \{b\} with the original support <3, 3> could become <2.5, 3.2> after adding Laplace noise. As we can see, the consistency constraint has been broken. The DP-topkP algorithm in [5] performs well in terms of data utility, while in this paper noise is reduced through combining the k-anonymity model with the differential privacy model.

Normally, there are two metrics when evaluating an algorithm in differential privacy: privacy and utility. Therefore, a synergy must be achieved. In [6], an approach to enhance data utility in differential privacy via microaggregation-based k-anonymity is proposed.

In this paper, microaggregated data is released by clustering the selected patterns based on the weight of
each pattern to achieve k-anonymity, and top-k frequent patterns are selected based on the microaggregated data through the exponential mechanism. Lastly, the true support of each frequent pattern is perturbed by adding Laplace noise. By adjusting the cluster size, a synergy between privacy and data utility can be achieved.

2. Background And Related Work

2.1 Mining Frequent Patterns

Frequent pattern mining is widely used in the area of knowledge discovery and can serve various purposes, among which mining association rules [7] and predicting user behavior [8] have been greatly studied.

Formally speaking, let \( t_i \) be a transaction record, and \( D = \{ t_1, t_2, \ldots, t_N \} \) be the transaction dataset that contains a few items. A set of items tending to appear together as a pattern shall be denoted as \( p_i \). In this paper, we apply the FP-Growth algorithm to the original dataset \( D \) to create a set of frequent patterns with support larger than the minimum support, denoted as \( P_{set} \). For each pattern \( p_i \in P_{set} \), assign a weight value \( w_i \) based on its frequency. The universe of denoted as \( P_{weight} \) stands for the weight value set of all the frequent patterns mined through the FP-Growth algorithm.

2.2 Differential Privacy

The differential privacy model for query operation in an interactive setting was originally proposed in [9]. The differential privacy model provides significantly robust and provable privacy preservation.

Definition 1: \( \varepsilon \)-differential privacy

As discussed in [1,10], for two given neighboring datasets \( D \) and \( D' \), a randomized mechanism \( A \) gives for any \( s \in \text{range} \,(A) \):

\[
Pr\,[A(D)=s] \leq e^{\varepsilon} \cdot Pr\,[A(D')=S].
\]

In this paper, the two neighboring datasets are or, where denotes a frequent pattern. Specifically, preserving differential privacy in a frequent pattern mining setting refers to the change in the probability distribution of the \( e^{\varepsilon} \), multiplicative - bounded by selecting a given frequent pattern or not.

2.3 Achieving k-anonymity based on Microaggregation

As shown by many researchers, the k-anonymity [11] model and its extensions have a significant influence on the research of privacy preservation. This model defines the attributes related to background knowledge as quasi-identifier attributes. Based on the same quasi-identifier, all records are generalized or suppressed so they can be separated into several equivalent groups. There are at least k records in each group. Although the k-anonymity model performs well in preventing identity disclosure, it cannot protect records against attribute disclosure.

For example, Table 1 shows a 3-anonymity dataset. Alice, Bob, and Carol share the same quasi-identifier attributes in Work (which defines whether the person works or not) and Age, and the AIDS attribute is what needs to be protected. Assume an intruder knows the target is in the group and Alice and Bob are not infected. He can therefore easily establish that Carol is infected. Several extensions of k-anonymity have been proposed such as l-diversity [12], t-closeness [13], \((\alpha, k)\) - anonymity [14], and M-invariance. However, these models all suffer from two vital defects. First, the privacy disclosure is related to the background knowledge of the intruder, and the background knowledge is difficult to fully define. Second, these approaches cannot be strictly proved.

However, the k-anonymity model is preferable to the differential privacy model from the perspective of data utility. Hence, as studied in [6], we preprocess the transaction dataset based on k-anonymity, and a microaggregation [15] algorithm is used as the anonymity algorithm with two procedures including clustering the records in the dataset based on the quasi-identifier attributes and replacing the original attributes with centroids, normally the mean.

3. Mining Frequent Patterns Through Microaggregation In Differential Privacy

In [6], the MDAV algorithm is applied to cluster the records in the dataset. For frequent pattern mining, each frequent pattern should be treated as a category rather than an arithmetic number. Therefore, the microaggregation operation should be applied to the weight of each frequent pattern, which is an arithmetic number. As discussed in section 2.1, we have achieved the frequent pattern set \( P_{set} \). Next, the weight of each pattern should be calculated based on the frequency in the exponential mechanism [16]. First, we define the ranking function \( Rank(D, P) \):

\[
Rank(D, P) = \sum_{i=1}^{\text{count}(t_i, p)} \text{count}(t_i, p), t_i \in D
\]

subject to \( \text{if } p \subset t_i \), \( \text{count}(t_i, p) = 1 \)

\text{else } \text{count}(t_i, p) = 0
\]

Since adding or deleting one transaction record will cause the support of a pattern to increase one or decrease one
at most, we know that the global sensitivity of the ranking function is $\Delta \text{Rank} = k$. With this in mind, the weight of each pattern should be depicted formally as:

$$w_i = \exp(\varepsilon \text{Rank}(D, p_i)) / \text{Rank}$$

where $w_i$ denotes the weight of pattern $p_i$ consisting of the weight set $P\text{weight}$.

### 3.1 Select top-k Frequent Patterns in the Exponential Mechanism

In this paper, the microaggregation operation is applied to $P\text{weight}$. (1) The records in $P\text{weight}$ are clustered into several groups with the decreasing weight of each pattern. Each cluster contains at least records. (2) All records in the individual cluster are replaced by a representative value, mostly the mean value within the cluster.

**Algorithm 1.** Release the differentially private data through microaggregation in the exponential mechanism.

let $P\text{weight}$ denote the set of weight value of all the frequent patterns

let $m$ be the minimal cluster size

set $i := 0$

while $|P\text{weight}| 2m$ do

$C_i \leftarrow m$ largest weight from $P\text{weight}$

$P\text{weight} : = P\text{weight} \setminus C_i$

$i := i + 1$

end while

$P\text{weight} \leftarrow$ replace each value $w_i \in P\text{weight}$ with the mean of the cluster

return $P\text{weight}$

After the microaggregation, we are able to obtain the set of microaggregated weight of each pattern $p_i$, denoted as $P\text{weight}$. The top-k frequent patterns should be selected from $P\text{set}$ based on the perturbed weight of each pattern.

**Algorithm 2.** Select top-k frequent patterns randomly.

For $i = 1$ to $|P\text{set}|$ do

set $Pr(P) = \frac{\sum_{i=1}^{\text{|Pset|}} w_i}{\text{|Pset|}}$

randomly select a pattern $p \in P\text{set}$

add $p$ to the top-k frequent pattern set $M\text{set}$

end for

### 3.2 Perturb the true support in the Laplace mechanism [10]

Since we selected top-k frequent patterns, to prevent support disclosure to the intruders, we need to add Laplace noise to the support of each pattern.

**Algorithm 3.** Perturb the support of each pattern.

let $support(p)$ return the support of $p$

for $i = 1$ to $|M\text{set}|$ do

$support(p) \leftarrow support(p) + \text{Lap}(k / \varepsilon_2)$

end for

return $M\text{set}$

### 3.3 Privacy analysis and utility analysis

In our approach, both the exponential mechanism and Laplace mechanism have been applied, respectively, to select top-k frequent patterns based on the masked weight and to perturb the true support of each pattern. In this section, we prove that the entire proposed mechanism preserves differential privacy.

**Proof 1.** Selecting top-k patterns preserves $\varepsilon_1$--differential privacy

In the counting scenario of mining frequent patterns, the global sensitivity of the ranking function could be denoted as $\Delta \text{Rank}$, and the probability of pattern $p$ being selected is formally depicted as:

$$Pr(p) = \frac{\exp(\varepsilon \text{Rank}(P\text{set}, p)) / 2 \Delta \text{Rank}}{\sum_{p' \in P\text{set}} \exp(\varepsilon \text{Rank}(P\text{set}, p')) / 2 \Delta \text{Rank}}$$

where $\varepsilon$ is the privacy budget [17]. We assign $\frac{\varepsilon_1}{k}$ to be the privacy budget to select an individual pattern $p_i$ and denote $P\text{set}'$ as the neighboring set of $P\text{set}$. Thus, the probability of one pattern being selected should be formally depicted as:

$$Pr(p_i) = \frac{\exp(\varepsilon_1 \text{Rank}(P\text{set}, p_i)) / 2k \Delta \text{Rank}}{\sum_{p' \in P\text{set}} \exp(\varepsilon_1 \text{Rank}(P\text{set}, p')) / 2 \Delta \text{Rank}}$$

Denote the selecting function by $\text{select}(P\text{set}, i)$, returning the pattern to be selected for the $i$th time. The differential privacy constraint for selecting an individual pattern can be proved as follows:

$$Pr(\text{select}(P\text{set}, i) = p) = \frac{\exp(\frac{\varepsilon_1}{2k}(\text{Rank}(P\text{set}, p) - \text{Rank}(P\text{set}', p)))}{\sum_{p' \in P\text{set}} \exp(\frac{\varepsilon_1}{2k}(\text{Rank}(P\text{set}, p') - \text{Rank}(P\text{set}', p')) / 2 \Delta \text{Rank})} \times \frac{\sum_{p' \in P\text{set}} \exp(\varepsilon_1 \text{Rank}(P\text{set}, p')) / 2 \Delta \text{Rank}}{\sum_{p' \in P\text{set}} \exp(\varepsilon_1 \text{Rank}(P\text{set}', p')) / 2 \Delta \text{Rank}}$$

$$\leq \exp\left(\frac{\varepsilon_1}{2k}\right) \times \exp\left(\frac{\varepsilon_1}{2k}\right) \times \frac{\sum_{p' \in P\text{set}} \exp(\varepsilon_1 \text{Rank}(P\text{set}', p')) / 2 \Delta \text{Rank}}{\sum_{p' \in P\text{set}} \exp(\varepsilon_1 \text{Rank}(P\text{set}, p') / 2 \Delta \text{Rank})}.$$
According to the definition of differential privacy, selecting one pattern preserves differential privacy with privacy budget $\frac{\varepsilon_1}{k}$. To select $k$ patterns, differential privacy is preserved with privacy budget $\varepsilon_1 \times k = \varepsilon$, according to sequential composition [18]. In total, $k$ patterns should be selected.

After applying the microaggregation operation to the weight of each pattern, the sensitivity of selecting the individual pattern is upper bounded by $\Delta \text{Sen}(Pset')/m$, where $\Delta \text{Sen}(Pset)$ denotes the sensitivity of the selecting function. As there are $n \times m$ clusters, the total sensitivity of the selecting function is upper bounded by $n \times m \times \text{Sen}(Pset)/m$. Thus, cluster size $m$ should be larger than $\sqrt{n}$ to reduce the magnitude of the exponential noise and improve data utility.

**Proof 2.** Perturbing the support of each pattern preserves $\varepsilon_1$—differential privacy

First, it should be proved that perturbing the support of an individual pattern $p$ preserves $\varepsilon_1/k$—differential privacy. Denote the set of the support of all patterns with perturbed support as $Sset'$, and denote the function to obtain the true support of pattern $p$ from dataset $D$ by $\text{tsup}(D, p)$ and the perturbed support by $\text{psup}(D, p)$. We define $\Delta$ as the global sensitivity of the perturbing function:

$$Pr(\text{psup}(D, p) \in Sset') = \frac{\exp(-\frac{\varepsilon_1|\text{tsup}(D, p) - \text{psup}(D, p)|}{k \Delta})}{\exp(-\frac{\varepsilon_1|\text{tsup}(D', p) - \text{psup}(D', p)|}{k \Delta})}$$

$$= \exp(-\frac{\varepsilon_1|\text{tsup}(D, p) - \text{psup}(D, p)| - |\text{tsup}(D', p) - \text{psup}(D', p)|}{k \Delta})$$

$$\leq \exp(-\frac{\varepsilon_1|\text{psup}(D, p) - \text{psup}(D', p)| - |\text{psup}(D, p) - \text{psup}(D, p)|}{k \Delta})$$

$$\leq \exp(-\frac{\varepsilon_1|\text{psup}(D, p) - \text{psup}(D', p)|}{k \Delta})$$

$$\leq \exp(-\frac{\varepsilon_1}{k \Delta})$$

As a result, differential privacy is preserved to perturb the support of an individual pattern. Again, differential privacy stands for the perturbation operation according to sequential composition with privacy budget $\frac{\varepsilon_2}{k} \times k = \varepsilon_2$.

In summary, the total privacy budget should be $\varepsilon = \varepsilon_1 + \varepsilon_2$, as the selecting operation and perturbation operation run sequentially in the algorithm.

### 4. Empirical Evaluation

#### 4.1 Evaluation Data and Metrics

To demonstrate the utility of the algorithm statistically, our approach has been applied to mine top-$k$ frequent patterns from retail market basket data [8], which contains 88,162 records and 16,470 items in total. At first, the FP-Growth algorithm is applied to mine all frequent patterns with support larger than the minimum support. In the experiment, we set the minimum support as 100, and $n = 6,445$ frequent patterns are mined.

Two utility metrics are used in the experiment: False Negative Rate (FNR) and Average Relative Error (ARE).

1. **FNR** measures the top-$k$ patterns that occur in $\text{topkP}(D)$ and not in $\text{topkP}(\tilde{D})$. $\text{topkP}(D)$ is the frequent pattern set mined from the original transaction dataset and $\text{topkP}(\tilde{D})$ is the perturbed dataset.

$$\text{FNR} = \frac{|\text{topkP}(D) \cup \text{topkP}(\tilde{D}) - \text{topkP}(\tilde{D})|}{k}$$

2. **ARE** measures the error introduced by the privacy preserving algorithm. $\text{trueSup}(p_i, \text{topkP}(D))$ is the true support of pattern $p_i$ in the selected top-$k$ patterns, and $\text{perSup}(p_i, \text{topkP}(D))$ is the perturbed support.

$$\text{ARE} = \frac{\sum_{p_i \in \text{topkP}(D)} |\text{trueSup}(p_i, \text{topkP}(D)) - \text{perSup}(p_i, \text{topkP}(D))|}{\text{trueSup}(p_i, \text{topkP}(D))}$$

![Diagram showing FNR and ARE metrics vs. epsilon values](image-url)
For the second group of experiments, $k = 100$ and $\varepsilon$ ranges from 0.5 to 1.5, while for the microaggregation operation the cluster size $m$ is set as $2 \times \sqrt{6445} \approx 160$.

5. Conclusion

When mining frequent patterns, the approach proposed here may be employed to enhance data utility while preserving differential privacy through microaggregation. Based on the exponential mechanism and the microaggregated weight of each pattern, the top-$k$ frequent patterns are selected. To enhance privacy, the true support of each selected pattern is perturbed by adding Laplace noise. In conclusion, a synergy between privacy and utility is achieved.

Acknowledgements

The research work was supported by National Natural Science Foundation of China under Grant No. 61272419, Prospective Research Project of Jiangsu Province under Grant No.BY2013095-3-02, No.BY2014089, No.BY2013039, No.BY2013037 and Lianyungang International Cooperation Project under Grant No. CH1304.

References


