PAPER

Privacy Preserving Association Rule Mining Revisited: Privacy Enhancement and Resources Efficiency[*]

Abedelaziz MOHAISEN[†a)], *Nonmember*, **Nam-Su JHO**[††b)], *Member*, **Dowon HONG**[††c)], *Nonmember*, *and* **DaeHun NYANG**[†††d)], *Member*

SUMMARY Privacy preserving association rule mining algorithms have been designed for discovering the relations between variables in data while maintaining the data privacy. In this article we revise one of the recently introduced schemes for association rule mining using fake transactions (FS). In particular, our analysis shows that the FS scheme has exhaustive storage and high computation requirements for guaranteeing a reasonable level of privacy. We introduce a realistic definition of privacy that benefits from the average case privacy and motivates the study of a weakness in the structure of FS by fake transactions filtering. In order to overcome this problem, we improve the FS scheme by presenting a hybrid scheme that considers both privacy and resources as two concurrent guidelines. Analytical and empirical results show the efficiency and applicability of our proposed scheme.

key words: *privacy preservation, association rule mining, data sharing, resources efficiency, performance evaluation*

1. Introduction

Data mining is a powerful tool for discovering knowledge, such as hidden predictive information, pattens and correlations, from large databases [1]. However, since the data itself may include information that lead to user identification, the privacy preserving data mining (PPDM) has emerged to become of a great interest [2]. In the PPDM settings, not only the accuracy of the data mining algorithms but also the privacy of data are considered essential [3]. Since the first work on PPDM by Agrawal et al. [2], several algorithms have been developed to treat the privacy in several settings. These algorithms are classified into *cryptographic* and *non-cryptographic* (randomization-based) algorithms [4]. The cryptographic algorithms for PPDM are shown to provide an accurate result of mining and provable privacy preservation guarantee [5], [6] at the expense of limited computational feasibility [6]–[9]. On the other hand, non-cryptographic approaches which utilize data randomization are shown to be computationally light though they provide low accuracy for achieving high privacy [7], [10]. In spite of their lack of provable guarantee to privacy preservation, the randomization-based algorithms have been favored over the cryptographic algorithms because of their computational feasibility merit. For that, several schemes are introduced in the literature directed to preserving privacy in data clustering [11]–[14], association rule mining [15]–[20], and data classification [21], [22], among mant others.

One of the interesting, though challenging, data mining applications is the association rule mining (ARM) [23], [24]. ARM is a well researched method for discovering *interesting relations* between variables in large databases. When adding the privacy concern to ARM, the privacy preserving association rule mining (PP-ARM) aims to discover such relations between the variables in data while maintaining its privacy. To do so, several algorithms have been introduced including those previously cited in [15]–[20], [25]. Among these works, Rizvi et al introduced MASK for PP-ARM [15]. In MASK (referred as PS), each bit in each transaction is altered into its binary complement with a probability p or kept as it is with a probability $1 - p$ (see Sect. 3 for details). Accordingly, the mining algorithm is modified so that an approximation of support and confidence are computed over the modified data given p. This algorithm has two advantages: (1) the privacy is quantified based on a sound mathematical definition and, (2) it does not require any memory overhead. However, its shortcoming is that the maximum achievable privacy is bounded (up to 0.89).

In another work (referred as FS), Lin et al used fake transactions to anonymize original data transactions for PP-ARM [18]. The FS has several advantages over other existing schemes. Particularly, (1) it uses an off-the-shelf mining algorithm and, (2) it provides a high theoretical privacy guarantee. However, its unmentioned drawback is the required high memory overhead for that privacy.

In this paper, we revise the FS scheme and show several results. (1) We explore an average-case notion of the privacy preservation in FS that expresses the real privacy attained. (2) We analyze the requirements of FS and show that, in order to provide a high privacy, the FS scheme requires an exhaustive amount of storage higher than that re-

Manuscript received June 1, 2009.
Manuscript revised October 19, 2009.

[†] The author is with University of Minnesota Twin Cities, Minneapolis, MN 55455, USA. This work was done while the author was working for ETRI.

[††] The authors are with Software and Contents Research Laboratory, Information Security Research Division, Cryptography Research Team, Electronics and Telecommunication Research Institute, Daejeon 700–305, Korea.

[†††] The author is with Inha University, Incheon 402–751, Korea.

[*] A short version of this paper was presented at the 9th Workshop on Information Security Applications (WISA 2008). The copyrights of the presented part are held by the authors. This work was supported by the IT R&D program of MKE/IITA. [2005-Y-001-05, Development of next generation security technology].

a) E-mail: mohaien@cs.umn.edu
b) E-mail: nsjho@etri.re.kr
c) E-mail: dwhong@etri.re.kr
d) E-mail: nyang@inha.ac.kr

DOI: 10.1587/transinf.E93.D.315

quired by PS [15] (see Sect. 5). (3) In practice, we show that the privacy provided by the FS can be breached given that the original transactions are not modified and kept in the released modified data. (4) To exploit the advantage of the FS and reduce its memory requirements, we introduce a hybrid scheme that utilizes both FS and PS schemes (see Sect. 6). (5) We introduce thorough theoretical and experimental analyses that demonstrate the achieved properties in both the revised and original schemes.

1.1 Why Does Privacy Matter?

The privacy concern rises particularly due to two contradicting goals from data which are utility and profit. In order to illustrate the importance of the privacy in the context of data mining, we provide several examples of real applications. The first application comes from the health-care area. In this application, a hospital would like to release patients' data for an external third party, who is typically not trusted, for research purposes. However, *insurance companies* (who act as an *attacker* on the private) have a great interest to know health record of the patients and their parents. Particularly, if a disease exist in the parents' record, it is highly probable that the children of this family will have the same disease. In order to maximize the profit, the insurance companies may increase the insurance on that family.

The second application is recalled from marketing and market analysis. In this example, a retailing company would like to know the pattern of customers' choice and future directions from a given marketing records that it already has. One of the possible options for that company is to outsource its own data to a third party that performs the mining task and discover any interesting patterns and provide them back to the company. While this data is not important for many people, it would be important for other companies competing in the same market (the *attacker*). Therefore it is required to provide an image of the data for required task without revealing additional information to the third party.

The third example shows how privacy research is being motivated by laws and regulations. According to several regulations and laws, personal data is must be preserved and cannot be stored permanently or used for making decision by any party. Particularly, since the ultimate goal of data mining algorithms is to build decisions on patterns driven from data, it is hard to eliminate the bias of decision based on gender or race. An example of such regulations includes HIPAA (Health Insurance Portability and Accountability Act) [26], PIPEDA (Canadian Personal Information Protection Act) [27], Directive 95/45/EC on the protection of personal data (of the European Union) [28] and ISO/TC 215 (an international standard) [29], among others.

1.2 Paper Organization

This paper is organized as follows. In Sect. 2 we introduce the preliminaries, definitions and notations. In Sect. 3 and Sect. 4 we summarize the PS and FS schemes for PP-ARM respectively. In Sect. 5, we revise the FS scheme showing its memory requirements for high privacy guarantee, average case privacy, fake transactions filtering, and further remarks for extension by comparing FS to PS. In Sect. 6 we introduce our hybrid scheme and its advantages over PS and FS apart in terms of privacy, resources, and error of mining result (both analytically and empirically). Concluding remarks are made in Sect. 8.

2. Preliminaries and Definitions

2.1 Major Notation

The major notations used through this paper are shown in Table 1. Also, other minor notations are defined where they are used through the rest of the paper.

2.2 Data Model

The market basket model is used for the ARM [†]. In this data model each user participates with a tuple (transaction) in the database where data tuples are with a fixed length and represented as sequences of 0's and 1's. The columns in the database represent the items whereas existence of 1 in a tuple indicates a purchase of the specified item and the existence of 0 indicates no purchase. Since users normally buy a smaller fraction of products than the whole number of products in the market, the number of 1's is much fewer than the number of 0's. The goal of the mining process is to compute the set of association rules in the database that meets a specific criterion. The data as a set of transactions $\mathcal{T}$ is represented as $\mathcal{T} = \{t_1, t_2, t_3, \ldots, t_N\}$ where $t_j \in \mathcal{T} = (a_1^{(j)}\, a_2^{(j)}\, a_3^{(j)}\, \ldots\, a_n^{(j)})$, $a_i^{(j)} \in t_j = 1$ if item i is purchased and $a_i^{(j)} \in t_j = 0$ otherwise [30].

2.3 Definitions

Definition 1. association rule [16]: Let the whole itemset be $\mathcal{I} = \{a_1, a_2, a_3 \ldots, a_n\}$ and T is a set of N transactions where $T = \{t_1, t_2, \ldots, t_N\}$ and each transaction t_i is a subset of $\mathcal{I}$. The association rule is a *statistical implication* which

Table 1 Notation.

Notation	Stands for
FS	PP-ARM algorithm using fake transactions in [18].
PS	PP-ARM algorithm using data masking in [15].
P_r^{PS}	reconstruction probability when using PS.
P_r^{FS}	reconstruction probability when using FS.
P_p^{PS}	quantification of preserved privacy in PS.
P_p^{FS}	quantification of preserved privacy in FS.
w	the ratio of fake to real transactions in FS.
R_1, R_0	reconstruction probability of 1's and 0's in PS respectively.
a	weight of 1's over 0's in PS scheme.
p	probability of altering bits to their complement in PS.

[†] This model is figurative and any ARM application can be exhibited according to this model.

can be expressed as $X \Rightarrow Y$ where $X, Y \subseteq \mathcal{I}$ and $X \cap Y = \phi$.

An association rule $X \Rightarrow Y$ is said to have a support s if $X \cup Y$ appears in $s\%$ of the transactions T [20]. Similarly, an association rule is said to have c confidence if $c\%$ of the T that satisfy X also satisfy Y. While the support is a measure of the significance of an association rule, the confidence is used as a measure of strength. An association rule is of interest if both c and s are greater than some threshold values. According to the *Apriori mining algorith*, finding the association rules in a dataset is equivlant to finding the frequent itemsets in that associations rule. An itemset is frequent if its *support* is greater than a threshold value.

Definition 2. Support of Itemset [18]: Let A be a set of n items where $\mathcal{I} = \{a_1, a_2, a_3 \ldots, a_n\}$ and T is a set of N transactions where $T = \{t_1, t_2, \ldots, t_N\}$ and each transaction t_i is a subset of I. The support of A is defined as follows:

$$\text{supp}^T(A) = \frac{\#\{t \in T | A \subseteq t\}}{N} \tag{1}$$

Example: Let the items be $\mathcal{I}$ ={m, c, p, b, j}, and the minimum support be $s_{min} = 3$. Also, let the set of transactions be $t_1 \sim t_8$ shown as follows

t_1={m, c, b} t_2 = {m, p, j} t_3 = {m, b}
t_4 = {c, j} t_5 = {m, p, b} t_6 = {m, c, b, j}
t_7 = {c, b, j} t_8 = {b, c}

From the transactions, we can systematically derive the representation matrix in terms of ones and zeros reflecting the existence or absence of an item in each transaction, respectively, as follows:

$$\begin{aligned}\mathcal{T} = [&(1\ \ 1\ \ 0\ \ 1\ \ 0),\ \ (1\ \ 0\ \ 1\ \ 0\ \ 1),\\ &(1\ \ 0\ \ 0\ \ 1\ \ 0),\ \ (0\ \ 1\ \ 0\ \ 0\ \ 1),\\ &(1\ \ 0\ \ 1\ \ 1\ \ 0),\ \ (1\ \ 1\ \ 0\ \ 1\ \ 1),\\ &(0\ \ 1\ \ 0\ \ 1\ \ 1),\ \ (0\ \ 1\ \ 0\ \ 1\ \ 0)]\end{aligned} \tag{2}$$

By applying the support model in (1) on $\mathcal{T}$, we obtain the frequent itemsets {m}, {c}, {b}, {j}, {m, b}, {c, b}, and {j, c} with supports $\frac{5}{8}, \frac{5}{8}, \frac{6}{8}, \frac{4}{8}, \frac{4}{8}, \frac{3}{8}$, and $\frac{3}{8}$ respectively.

Definition 3. Privacy measure [18]: The privacy is defined as the probability according to which the distorted data can be reconstructed.

Definition 4. False positive σ^+: An error that happens when $k-$itemset with a support slightly less than s_{min} is supported with more transactions than other $k-$itemsets in the disguised data (i.e., these transactions are included in the counting though not being frequent). Let R and F be the reconstructed and real sets of frequent itemset, σ^+ is then defined as $\sigma^+ = \frac{|R-F|}{|F|}$

Definition 5. False negative σ^-: An error that happens when $k-$itemset with a support slightly greater than or equal s_{min} is supported with less transactions than other $k-$itemsets. In this scenario, these $k-$itemsets are not counted as frequent though they are frequent. Similar to σ^+, σ^- is defined as $\sigma^+ = \frac{|F-R|}{|F|}$.

Definition 6. Theoretical privacy: privacy measured using straightforward mathematical formulation and consider only an attacker who tries to find real transactions (fully) at random without any further knowledge

Definition 7. Real privacy: privacy achieved in practice when considering other circumstances beside random selection. E.g., an attacker has some knowledge about the pattern of choice in the dataset, the attacker may apply filtering, etc.

3. MASK for PP-ARM

In this section we overview the distortion procedure of set of transactions $\mathcal{T}$ represented according to the description in Sect. 2.2. To preserve the privacy, the data owner performs the following two steps

- Each tuple in the database is considered as a random variable $X = \{X_i\}$ where $X_i = 0$ or 1.
- The distortion follows the following procedure: $Y = \text{distort}(X)$ where $Y_i = X_i \oplus \bar{r}_i$ where $\bar{r}_i$ is complement of r_i which is a realization of a random variable with the probability distribution function $f(r) = \mathsf{bernoulli}(p)$ for $0 \leq p \leq 1$.

The technical implication of the randomization process is that r_i takes a value 1 with a probability p and 0 with a probability $1-p$. When $r_i = 1$, the original bit X_i in the data tuple is kept same and when $r_i = 0$ the original bit X_i is altered to its complement. The privacy of the PS scheme is estimated by the probability according to which the reconstruction of zeros and ones is possible

1. Reconstruction of ones according to $R_1 = P_r\{Y_i = 1|X_i = 1\}P_r\{X_i = 1|Y_i = 1\} + P_r\{Y_i = 0|X_i = 1\}P_r\{X_i = 1|Y_i = 0\} = \frac{s_0 \times p^2}{s_0 \times p + (1-s_0) \times (1-p)} + \frac{s_0 \times (1-p)^2}{s_0 \times (1-p) + (1-s_0) \times p}$. Note that s_0 here is the average support of an item in the database. For a general form where supports of items are different, please see [15].
2. Reconstruction of zeros according to $R_0 = P_r\{Y_i = 1|X_i = 0\}P_r\{X_i = 0|Y_i = 1\} + P_r\{Y_i = 0|X_i = 0\}P_r\{X_i = 0|Y_i = 0\} = \frac{(1-s_0) \times p^2}{(1-s_0) \times p + s_0 \times (1-p)} + \frac{(1-s_0) \times (1-p)^2}{s_0 \times p + (1-s_0) \times (1-p)}$.

The probabilities in 1 and 2 capture the round trip probability by moving from the original dataset to the randomized dataset (i.e., randomization) and moving back from the randomized dataset to the original dataset (i.e., reconstruction) for each bit [15]. From 1 and 2, the overall probability of successful reconstruction of 1's and 0's is $P_r^{\mathsf{PS}} = aR_1 + (1-a)R_0$ where a is a privacy parameter given to weight 1's over 0's [15]. The privacy attained in the PS is $P_p^{\mathsf{PS}} = 1 - P_r^{\mathsf{PS}} = 1 - (aR_1 + (1-a)R_0)$.

At the miner side, in order to compute the support for a k-itemset, the following generic form is used

$$C^X = M^{-1}C^Y \tag{3}$$

where C^Y is the column vector defined as

$$C^Y = [c^Y_{2^k-1}, \ldots, c^Y_1, c^Y_0], \tag{4}$$

in which c^Y_k is the count of the tuple in Y that has the linear form k and C^X is defined as

$$C^X = [c^X_{2^k-1}, \dots, c^X_1, c^X_0], \tag{5}$$

where c^X_k is the count of the tuple in X that has the linear form k, where X is the original data and Y is the distorted data. M is a $(2^k - 1) \times (2^k - 1)$ with entries $m_{i,j}$ defined as the probability that a tuple of the binary form c^X_i in X goes to the tuple of the form c^Y_j in Y. For instance, in transactions with 2 items, $m_{2,3}$ is the probability that (1 0) in X is mapped to (1 1) in Y which is $p(1 - p)$.

4. PP-ARM Using Fake Transactions

The PP-ARM using fake transactions (FS for brevity) is introduced in by Lin et al in [18]. FS uses fake transactions as noise in between of real transactions in the dataset. The privacy in FS is determined by the *quality* and *quantity* of the fake transactions. The *quantity* of fake transactions is determined according to the parameter w which stands for the ratio of fake transactions to real transactions and the parameter l which stands for the average length of a single fake transaction. The parameter l is chosen to be same as the average length of the real transactions and the parameter w is chosen based on the desirable privacy to be attained (P^{FS}_p). Let the hardness of filtering the real transactions from the fake transactions, P^{FS}_r, be expressed as

$$P^{\mathsf{FS}}_r = \frac{N}{N + Nw} = \frac{1}{1 + w}. \tag{6}$$

Then, P^{FS}_p is, by definition, determined as $P^{\mathsf{FS}}_p = 1 - P^{\mathsf{FS}}_r = 1 - \frac{1}{w+1}$.

The FS scheme consists of two phases which are data anonymization phase and data mining phase. The data anonymization phase consists of the following:

1. Determine l_i as a realization of uniformly distributed random variable (UDRV) with mean l equal to the average length of the real transactions (i.e., $1 \leq l_i \leq 2l - 1$).
2. Determine $w^{(i)}$ as the number of fake transactions to be inserted between two real transactions with index i and index $i + 1$ in the dataset. For a predefined w, $w^{(i)}$ is determined as a realization of a UDRV with mean w (i.e., $1 \leq w_i \leq 2w - 1$).
3. l_i number of items are selected from $\mathcal{I}$ to construct a fake transaction.
4. The process is performed for $w^{(i)}$ times for the current insertion.
5. The $w^{(i)}$ number of fake transactions generated in steps 3 and 4 are inserted in between of the real transactions with indexes i and $i + 1$.

The procedure in steps 1 through 5 is performed for the whole set of pairs of tuples in the database (i.e., $N - 1$ pairs).

To learn the association rules from the anonymized data, the data mining phase is performed as follows.

1. The new minimum support for a transaction of k-itemset in the anonymized transactions T' is computed.
2. Using any off-the-shelf algorithm, the association rules are driven according to the new minimum support.

The procedure of computing the new minimum support is as follows. Given a fake transaction t of length l_t and k-itemset A, the probability that t supports A is

$$p_k = \frac{C^{n-k}_{l_t-k}}{C^n_{l_t}} = \frac{C^{l_t}_k}{C^n_k}, \text{(when } l_t \geq k \text{ and 0 otherwise).} \tag{7}$$

Note that $C^{l_t}_k$ stand for the overall number of itemsets in t with length k. The number of fake transactions that support k-itemset is approximately

$$\sum_{l_t=k}^{2l-1} \frac{C^{l_t}_k}{C^n_k} \times \frac{w \times N}{2l - 1} = \frac{wN}{C^n_k(2l - 1)} \sum_{l_t=k}^{2l-1} C^{l_t}_k. \tag{8}$$

Note that the approximation in Eq. (8) assumes that every length is support with $\frac{w \times N}{2l-1}$ fake transactions in the disguised data at average. This approximation is the main reason for the resulting error (σ^+ and $\sigma-$) as some lengths can be supported with more or less than the average number. Assume the support of $A \in T'$ is s' (i.e., $\text{supp}^{T'}(A) = s'$), then the number of transactions in T' that support A is $s'(1 + w)N$. Therefore, the number of real transactions that support A in T' is

$$s'(1 + w)N - \frac{wN}{C^n_k(2l - 1)} \sum_{l_t=k}^{2l-1} C^{l_t}_k. \tag{9}$$

Let the real support be s, then we can write Eq. (9) as $s = s'(1 + w) - \frac{w}{C^n_k(2l-1)} \sum_{l_t=k}^{2l-1} C^{l_t}_k$. Therefore, the new minimum support s' is driven as

$$s'_k = \frac{s_{min} + \frac{w}{C^n_k(2l-1)} \sum_{l_t=k}^{2l-1} C^{l_t}_k}{1 + w} \tag{10}$$

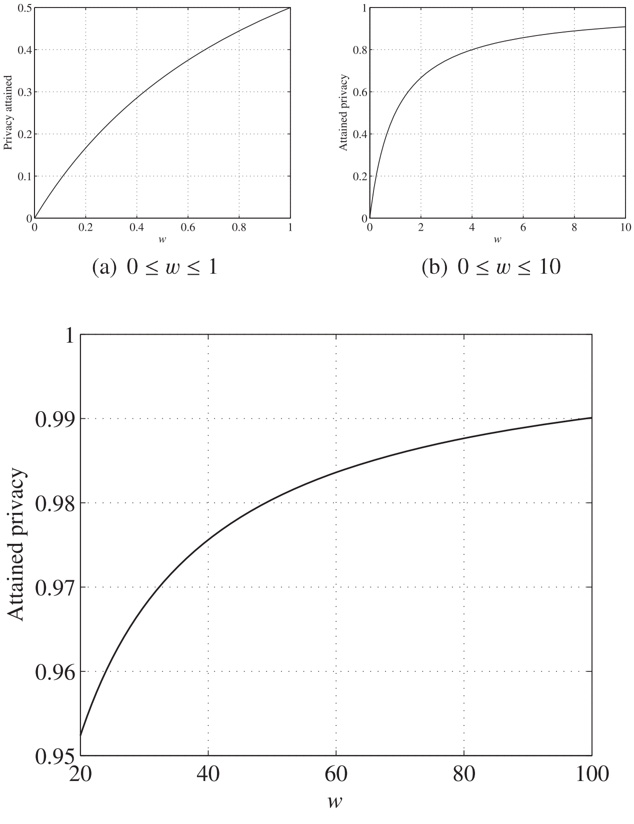
Since all parameters in the right-hand side of Eq. (10) are known, we can learn the association rules in T' given only the minimum support s_{min} in T. For further details on the FS scheme and its optimization, refer to [18].

5. PP-ARM Revisited

In this section, we revisit the FS scheme and introduce three main results which are: (i) we show that the FS scheme is resources exhaustive in terms of high memory in order to provide a reasonable level of privacy, (ii) we show that the theoretical quantification of the privacy in the FS follows the worst-case study while the average-case can be more realistic descriptor for the privacy attained, and (iii) we show that using two round attack where the first attack is done by applying common filters on the data and the second by the random selection, we show that the privacy can be less than the above two cases.

5.1 Requirements Analysis of the FS Scheme

The privacy of the FS scheme depends on the parameters l and w which both determine the quality and quantity of fake transactions. While l does not have *any* effect on the required memory since each transaction has a fixed length, w, which is the determining factor of the privacy (as shown in (6)), has a great effect. The privacy attained by FS is defined as $P_p^{\mathsf{FS}} = 1 - P_r^{\mathsf{FS}} = 1 - \frac{1}{w+1}$. In order to attain a relatively high privacy, w need to be large enough. For instance, to achieve a privacy of 90% (0.9 on the 1-scale), w needs to be at least 11. That is, the required additional memory for representing and storing the fake transactions in $T^{'}$ will be 11 times the original database size. To illustrate the growth of such functions, Fig. 1 shows different growth regions of w. In Fig. 1 (a), the growth is shown for $0 \leq w \leq 1$ which reflexes the fast growth region attaining 0.5 privacy. Figure 1 (b), shows the range of $0 \leq w \leq 10$ from which we observe that an increment of 9 in w leads to only 0.4 additional privacy preservation over the case $w = 1$. Finally, for $10 \leq w \leq 100$, Fig. 1 (c) shows that the variation of w by 90 would add a preserve the privacy 0.04 more than the case of $w = 10$ to accumulate 0.99 for $w = 100$. This problem is particularly critical when we aim to achieve a high privacy. On the other hand, the required computation linearly depends on the size of the dataset in which the association rules to be learned. That is, the increment of the database size in $T^{'}$ will require w times computational power more than the computation required for the association rules discovery in T only.



(a) $0 \leq w \leq 1$

(b) $0 \leq w \leq 10$

(c) $20 \leq w \leq 100$

Fig. 1 The attained privacy versus the required w that reflex the required overhead in terms of memory and computation.

5.2 Average-Case for Privacy Quantification

The privacy attained in the FS scheme according to the description in [18] and summarized in Sect. 4 is indicated as the worst-case privacy. The worst privacy is driven by assuming that the reconstruction probability of any tuple in the anonymized database $T^{'}$ is equal to the reconstruction probability of the first tuple. In other words, the probability of all tuples is assumed to be equal. However, since the attacker is assumed to reconstruct tuples successively without replacement, there is a necessity for defining an average case privacy that considers the privacy attained at any time through the life of the data. In the following (Claim 1), we define the average-case privacy and show its relation to the worst-case privacy in [18].

Claim 1. average-case privacy The quantification of privacy in [18] considers the best reconstruction probability of a single record (i.e., worst case privacy measure) while the real privacy preserved (at average) is greater than the worst case quantification.

Proof. Let an adversary $\mathcal{A}$ interested in recovering the whole set of *real transactions* by applying a random selection process. For the sequence of trials to obtain the transactions $t_1 \ldots t_N \in T^{'}$, the following is the probability for successful reconstruction of the N real transactions anonymized in the set of $w \times N$ fake transactions.

$$\begin{aligned} P_r = \frac{1}{N}\Bigg[& \frac{N}{wN+N} + \frac{N-1}{wN+N-1} + \ldots \\ & + \frac{N-(N-1)}{wN+N-(N-1)}\Bigg] \\ = & \frac{1}{N} \times (p_0 + p_1 + \cdots + p_{N-1}) \end{aligned} \tag{11}$$

Then, we verify that $p_i > p_{i+1}$ for $1 < i < N-1$. Let $i = 1$ then $\frac{N}{wN+N} > \frac{N-1}{wN+N-1}$. By multiplying both sides by $\frac{wN+N-1}{N}$, we get that $\frac{wN+N-1}{wN+N} > \frac{N-1}{N}$ which is valid for any $w > 0$ and $N > 2$ (note that both conditions are always rationally satisfied under the real data assumptions). We can similarly extend the above result to $i > 1$ and state that $c \times p_i > \sum_{j=0}^{c} p_{i+j}$ for any $i \geq 1$ and $c \geq 1$. That is, by substituting $i = 1$ and $j = N-1$, we get that $N \times p_1 > \sum_{i=0}^{N-1} p_i$ which means $p_1 > \frac{1}{N}\sum_{i=0}^{N-1} p_i$. However, $\frac{1}{N}\sum_{i=0}^{N-1} p_i = P_r$ and $p_1 = P_r^{\mathsf{FS}}$. Therefore, $P_r^{\mathsf{FS}} > P_r$. From this final result we get that $1 - P_r^{\mathsf{FS}} < 1 - P_r$ which means that $P_p^{\mathsf{FS}} < P_p^{\mathsf{FS}'}$ where P_p^{FS} and $P_p^{\mathsf{FS}'}$ are the quantification of privacy preserved in the FS scheme in [18] and the average case introduced by us, respectively. □

Note that the last result of the average-case privacy

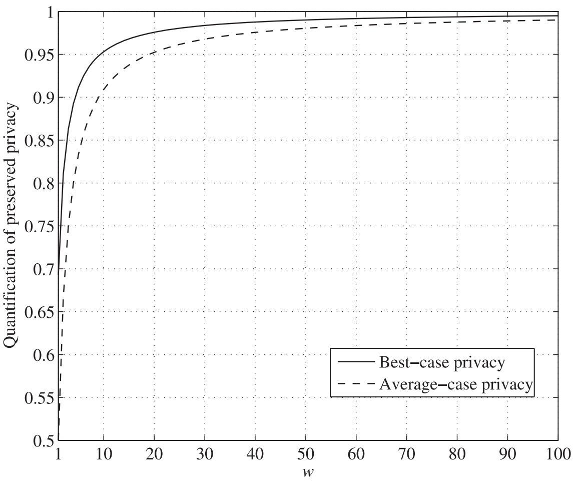


Fig. 2 The average versus the worst case privacy preservation.

quantification is more general in expressing the attained privacy according to the privacy preservation procedure introduced in [18]. Also, this measure is more suitable for observing the attack below. A comparison between the average case privacy and the worst privacy is shown in Fig. 2.

5.3 On fake Transactions Filtering

The main concern in [18] has been the filtering and reconstruction of *real transactions* inserted in between of the fake transactions. However, an adversary $\mathcal{A}$ might be interested in removing some of the fake transactions which are obvious in order to maximize the chances of obtaining the real transactions in the remaining set of transactions according to the aforementioned privacy quantification model.

The above scenario is possible because it is hard, if not impossible, to generate fake transaction that are typically indistinguishable from the real transactions. This is particularly obvious when the distribution of the dataset is unknown or biased. In the settings of FS, this shortcoming introduce a great chance for filtering *weak fake transactions* using many off-the-shelf statistical filters. Moreover, given additional information on the distribution of the user choice in the original data, it is further possible to filter high amount of fake transactions. Generally, the filtering may take one, or even both, of the following forms:

1. **Random filtering:** since the number of the fake transactions in $T^{'}$ is greater than the number of real transactions, specially when $w > 1$, it is *more likely* to select a transaction at random such that the selected transaction belongs to the set of fake transactions.
2. **Guided filtering:** given enough information to $\mathcal{A}$ about the distribution of the real transactions representing the choice of users, $\mathcal{A}$ can with a high certainty filter a large amount of the fake transactions.

Let the efficiency of the filter applied on $T^{'}$ be γ where $0 \leq \gamma \leq 1$. Then, the model in Eq. (11) can be extended to contain the filtering impact as:

$$P_r^{(\gamma)} = \frac{1}{N}\left[\frac{N}{(1-\gamma)wN+N} + \frac{N-1}{(1-\gamma)wN+N-1} + \dots \right. \\ \left. + \frac{N-(N-1)}{(1-\gamma)wN+N-(N-1)}\right]. \qquad (12)$$

Similarly, we derive the average-case privacy as

$$P_p^{\mathsf{FS}'(\gamma)} = 1 - P_r^{(\gamma)} = 1 - \sum_{i=0}^{N-1} \frac{N-i}{(1-\gamma)wN+N-i}. \qquad (13)$$

To illustrate the impact of the filtering on the privacy preservation, Table 2 and Table 3 show the quantified privacy preservation for different filtering efficiency parameters γ according to different values of w.

Table 2 Quantified privacy preservation under several filtering efficiency factors ($\gamma = 0.0$ to $\gamma = 0.7$) and for $w = 1$ to $w = 5$.

	$w = 1$	$w = 2$	$w = 3$	$w = 4$	$w = 5$
$\gamma = 0.0$	0.6929	0.8108	0.8629	0.8925	0.9115
$\gamma = 0.1$	0.6722	0.7951	0.8506	0.8823	0.9029
$\gamma = 0.2$	0.6485	0.7766	0.8358	0.8701	0.8925
$\gamma = 0.3$	0.6208	0.7544	0.8177	0.8549	0.8795
$\gamma = 0.4$	0.5882	0.7271	0.7951	0.8358	0.8629
$\gamma = 0.5$	0.5490	0.6929	0.7660	0.8108	0.8410
$\gamma = 0.6$	0.5007	0.6485	0.7271	0.7766	0.8108
$\gamma = 0.7$	0.4395	0.5882	0.6722	0.7271	0.7660

Table 3 Quantified privacy preservation under several filtering efficiency factors ($\gamma = 0.0$ to $\gamma = 0.7$) and for $w = 6$ to $w = 10$.

	$w = 6$	$w = 7$	$w = 8$	$w = 9$	$w = 10$
$\gamma = 0.0$	0.9248	0.9347	0.9422	0.9482	0.9531
$\gamma = 0.1$	0.9174	0.9281	0.9363	0.9429	0.9482
$\gamma = 0.2$	0.9083	0.9200	0.9291	0.9363	0.9422
$\gamma = 0.3$	0.8969	0.9099	0.9200	0.9281	0.9347
$\gamma = 0.4$	0.8823	0.8969	0.9083	0.9174	0.9248
$\gamma = 0.5$	0.8629	0.8795	0.8925	0.9029	0.9115
$\gamma = 0.6$	0.8358	0.8549	0.8701	0.8823	0.8925
$\gamma = 0.7$	0.7951	0.8177	0.8358	0.8506	0.8629

5.4 Remarks and Extensions

We compare the FS versus PS schemes and point out their strength and shortcomings. The PS scheme requires no memory overhead except from that required for representing the data itself while the FS scheme requires memory space for the additional wN number of fake transactions used to disguise the real transactions. Such memory can be tens of gigabytes for large datasets limiting the later schemes' feasibility and applicability.

The PS scheme has an upper bound for the quantified privacy. That is, for the maximum possible p, the attained privacy is equal to 89% [15]. This quantified privacy is possibly sufficient for some applications but can be a great breach for privacy-critical applications [17]. On the otherh and, the privacy resulting from the FS scheme is merely dependent upon the allowed amount of overhead.

Both schemes' excessive privacy lead to a relatively higher error of the mining algorithm. Also, while the PS scheme requires modification in the mining algorithm

Table 4 Comparison between the FS and PS schemes.

Feature	PS scheme	FS scheme
Memory Overhead	0	$O(wN)$.
Computation	$\sim N$	$\sim wN$
Mining Algorithm	Modified	off-the-shelf

to maintain a reasonable computation overhead, the FS scheme can use any off-the-shelf algorithm for mining. Table 4 summarizes a concluding comparison between the two schemes.

6. Hybrid Scheme for Association Rules

The FS scheme introduces some great properties at the expense of drawbacks which are summarized as follows: (1) The FS scheme introduces theoretically high privacy at the expense of high resources in term of memory and computation and, (2) The FS uses any off-the-shelf mining algorithm though the presence of the bare real transactions within the disguised data enables fake transactions filtering that leads to reducing the originally attained privacy. Based on that, there is a great chance to utilize and extended FS scheme that maintains the advantages and mitigates the drawbacks. Here, we recall the PS explained in Sect. 3 and explain how a hybrid scheme of both the PS and FS (referred as HS) will support the aforementioned goals. Our scheme utilizes the two introduced schemes above to have their advantages together and reduce from their disadvantages specially related to the memory overhead and limited privacy.

6.1 HS for PP-ARM

Our hybrid scheme (HS for brevity) works as follows: first fake transactions are produced using the same way of the FS scheme and inserted in between of the real transactions for the whole set of transactions in the database then the modified database is distorted using the procedure of the PS scheme. The details of the data distortion part of the scheme are shown in Fig. 3.

On the other hand, the mining algorithm for checking whether a k-itemset is frequent or not is made as a combination of both schemes. Figure 4 shows the procedure of the mining algorithm.

To study the characteristics of the HS scheme, we use the following three criteria (1) Privacy measure (Lemma 1), (2) Error measure, (3) Overhead measures in terms of computation and memory (Lemma 2).

6.2 Privacy of HS

By definition, the successful reconstruction probability of a single tuple in the HS scheme is given as the probability that a selected tuple from the disguised data at random belongs to the original dataset and that each bit in this tuple is reconstructed correctly. That is, reconstruction probability in HS is defined as

```
Input: Dataset of $N$ transactions $T = \{t_1, t_2, \ldots, t_N\}$, $p$, $a$.
Output: disguised transactions $T^{'}$.

1. Compute $l = \frac{1}{N}\sum_{j=0}^{N} l_j$.
2. Let initial $i = 1$
3. Generate $w_i$ as a realization of a UDRV with $\mu = w$ and $l_i$ as a
   realization of a UDRV with $\mu = l$.
4. Generate a fake transaction of length $l_i$; repeat that for $w_i$ times.
5. Insert the set of $w_i$ fake transactions in between of the two real
   transactions with index $i, i + 1$.
6. Set $i = i+1$, repeat steps 3 through 5 till $i = N$. The accumulated
   dataset (including fake transactions) are donated as $T^{'}$.
7. Shuffle the order of transactions in $T^{'}$ to avoid obvious filtering
   when $w \leq 1$.
8. For each transaction $t_i \in T^{'}$, generate a transaction $t_i^{'}$ of $n$ bits
   made as realizations of Bernoulli random variable of probability $p$.
   Compute $t_i^{''} = t_i \oplus t_i^{'}$ where $\oplus$ is a bitwise exclusive or
   operation.
9. Release the disguised dataset as $T^{''} = \{t_1^{''}, t_2^{''}, \ldots, t_n^{''}\}$
```

Fig. 3 The data disguising process of the HS scheme. UDRV stands for uniformly distributed random variable with a mean μ.

```
Input: Distorted dataset $T^{''} = \{t_1^{''}, t_2^{''}, \ldots, t_N^{''}\}$, $p$, $s_{min}$
Output: Determine whether an itemset is frequent or not.

1. Given $s_{min}$, compute $s^{'}$ according to Eq. (10).
2. Compute the count of candidates in $Y$ according to Eq. (4).
3. Compute the matrix $M$ according to the description in section 3.
4. Estimate the counts of the $k$-itemset candidates (i.e., in Eq. (5))
   by applying Eq. (3).
5. Compute the support $s_k$ of the $k$-itemset by applying Eq. (1).
6. If the $s_k \geq s^{'}$, the $k$-itemset is frequent, otherwise it is not.
```

Fig. 4 Itemset discovery in the HS scheme. This algorithm is performed for every k-itemset to determine if it is frequent or not.

$$P_r^{\mathsf{HS}} \triangleq P_r^{\mathsf{FS}} P_r^{\mathsf{PS}} = \frac{P_r^{\mathsf{PS}}}{1+w}. \tag{14}$$

Accordingly, we define the attained privacy as

$$P_p^{\mathsf{HS}} \triangleq 1 - P_r^{\mathsf{HS}} = 1 - \frac{P_r^{\mathsf{PS}}}{1+w}. \tag{15}$$

Particularly, since both P_r^{FS} and P_r^{PS} are less than one, the resulting probability P_r^{HS} is always less than the smallest of them. That is, $1 - P_r^{\mathsf{FS}} > \max\{1 - P_r^{\mathsf{FS}}, 1 - P_r^{\mathsf{PS}}\}$ which means that $P_p^{\mathsf{HS}} > \max(P_p^{\mathsf{FS}}, P_p^{\mathsf{PS}})$ by definition.

Lemma 1: The quantified privacy preserved using our hybrid scheme HS is higher than the privacy preserved using PS or FS alone.

Proof (sketch). Given that $0 \leq P_r^{\mathsf{FS}} \leq 1$ and $0 \leq P_r^{\mathsf{PS}} \leq 1$ then it is straightforward to show that $P_r^{\mathsf{FS}} P_r^{\mathsf{PS}} \leq P_r^{\mathsf{FS}}$ and $P_r^{\mathsf{FS}} P_r^{\mathsf{PS}} \leq P_r^{\mathsf{PS}}$. That is, $1 - P_r^{\mathsf{FS}} P_r^{\mathsf{PS}} \geq 1 - P_r^{\mathsf{FS}}$ and $1 - P_r^{\mathsf{FS}} P_r^{\mathsf{PS}} \geq 1 - P_r^{\mathsf{PS}}$ which yield $P_p^{\mathsf{HS}} \geq P_p^{\mathsf{FS}}$ and $P_p^{\mathsf{HS}} \geq P_p^{\mathsf{PS}}$ respectively. □

As a special case, it can be easily shown that our schemes' attained privacy is higher than PS scheme when P_p^{PS} equals to its maximum value (i.e., minimum P_r^{PS}).

Beside the attained theoretical privacy which is shown to be higher than the in the FS scheme, the HS scheme provides better resistance to fake transactions filtering than FS scheme in practice (i.e., results in a higher real privacy than the one estimated theoretically). Given that the order of transactions is shuffled and the set of fake transactions along with the real transaction is disguised, chances for applying meaningful filtering based on patterns are very low. For instance, one possible method of filtering is to choose transactions at random and assert that they are real transactions. However, even if the attacker succeed in distinguishing the real transactions, the attacker will have to go through the reconstruction procedure of the PS scheme with low success probability. On the contrary, once the attacker succeed in the first stage of the attack on FS, all filtered transactions are immediately used to breach the privacy.

6.3 Resources Consumption of HS

Unlike the FS scheme, the HS scheme requires less memory for the same level of privacy given that $p > 0$.

Lemma 2: For same privacy level, our HS scheme requires less storage than FS scheme.
Proof. Let w_1 and w_2 be two parameters defined for FS and HS schemes respectively. The privacy attained by each scheme is given as $P_p^{\mathsf{FS}} = 1 - \frac{1}{1+w_1}$ and $P_p^{\mathsf{HS}} = 1 - \frac{P_r^{\mathsf{PS}}}{1+w_2}$. By setting $P_p^{\mathsf{FS}} = P_p^{\mathsf{HS}}$ (i.e., attained privacy is equal in both schemes) we get that:

$$P_r^{\mathsf{PS}} = \frac{1+w_2}{1+w_1} \tag{16}$$

However since P_r^{PS} is less than 1 (more specifically, maximum P_r^{PS} is equal to 0.89), the above equality is only possible when $w_2 \leq w_1$. □

Example: to attain a privacy $P_p^{\mathsf{FS}} = P_p^{\mathsf{HS}} = 0.95$ when $P_r^{\mathsf{PS}} = 0.3$, it is enough to set $w_2 = 5$ while w_1 must be at least 19.

6.4 Error Measurement

To study the impact of using both schemes in one hybrid scheme on the resulting error represented by the false negative and false positive in Definition 4 and Definition 5 we developed and used the ppAR-Discovery. The ppAR-Discovery incorporates the Apriori algorithm for association rules discovery [31], the modified Apriori algorithm for discovering association rules in MASK [15], the distortion scenario of FS shown in Sect. 4, the distortion part of PS scheme explained in Sect. 3, and the hybrid scheme described in Fig. 3 and Fig. 4.

We conduct the experiment to evaluate the error rate on the dataset BMS-WebView-1 [24]. The used dataset consists of 59,602 transactions where each transaction consists of 497 items and the length of transaction at average (i.e., l) is equal to 2. We further set w with two values: 2 and 4 to generate fake transactions according to the procedure in 4 and set $p = 0.499$ according to which the privacy of PS scheme is determined. The measurements for the error is shown in Table 5 for different minimum support values.

Table 5 Error of mining in terms of false positive σ^+ and false negative σ^- for HS versus FS considering different parameters w and for $p = 0.5$ and different minimum support values.

		$s_{min} = 0.005$		$s_{min} = 0.0025$		$s_{min} = 0.001$	
scheme	w (privacy)	σ^+	σ^-	σ^+	σ^-	σ^+	σ^-
HS	2 (0.833)	4.013	2.728	2.341	2.340	2.172	1.503
FS	2 (0.667)	2.985	1.493	1.607	1.607	1.102	0.701
HS	4 (0.900)	6.731	4.275	4.762	3.698	1.591	1.620
FS	4 (0.800)	4.975	2.985	3.214	2.501	1.027	1.152

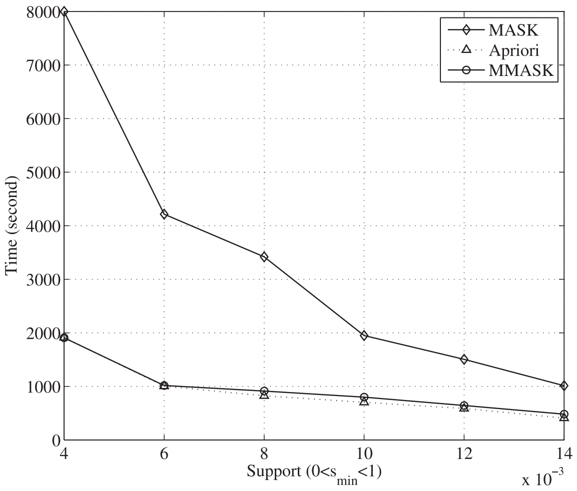


Fig. 5 Computation overhead (without privacy consideration).

Note that the correctness of the mining results depend on both of the desired privacy to be attained and the minimum support s_{min}. Particularly, higher percents of errors (both σ^+ and σ^-) occur when higher privacy is demanded and higher minimum support is assigned (as shown in Table 5). The intuition beyond the increase of error is that when larger number of fake transactions are added to the real transactions, it is more likely that some itemsets with some lengths will have more or less support by fake transactions than other itemsets.

6.5 Computation and Memory Comparison

In order to measure the computation requirements of HS, PS, and FS schemes, we use our implementation of ppARM-Discovery and the IBM Almaden dataset T10I8D100k which contains 100,000 transactions each of which has 10 items at average [32]. T10I8D100k requires 4.4 Megabytes of storage. We measure the time required for discovery association rules that dataset for different support values and using different algorithms. Namely, we use the Apriori algorithm, MMASK (modified version of Mask) and MASK. First, we apply each of these algorithms on the dataset without privacy consideration, where possible, and the time measurements are shown in Fig. 5 (MMASK and MASK are applied on disguised data with parameters used to for Table 6). We observe that the running time of MASK scheme (PS) is greater in magnitude than that required for Apriori and MMASK for small supports in particular. Because the Apriori algorithm does not use any un-

Table 6 A comparison of the running time between several association rule mining algorithms using several PP-ARM schemes for different w values. Shaded rows are measurements for HS scheme, unshaded rows are readings of FS scheme. PS occurs in the HS scheme when $w = 0$ and s_{min}, ARM without privacy consideration occurs in PS scheme when $w = 0$.

s_{min}	$w = 0$	$w = 1$	$w = 2$	$w = 3$	$w = 4$	$w = 5$
0.004	1902.538	3810.154	5707.420	7614.436	9518.382	11416.41
0.004	1915.518	3840.444	5757.377	7666.508	9565.712	11493.32
0.006	1006.857	2030.639	3020.090	4036.387	5031.731	6044.294
0.006	1020.402	2030.883	3059.891	4072.110	5079.988	6112.374
0.008	0819.317	1644.548	2457.953	3284.580	4092.814	4930.341
0.008	0918.858	1839.837	2760.474	3683.245	4604.156	5503.679
0.010	0710.765	1415.095	2129.118	2848.841	3550.869	4261.084
0.010	0793.155	1588.689	2380.345	3178.311	3960.590	4755.187
0.012	0583.790	1171.383	1762.320	2335.563	2904.199	3508.973
0.012	0643.195	1276.313	1923.231	2564.564	3219.250	3847.313
0.014	0412.355	0814.620	1218.326	1656.193	2059.438	2482.124
0.014	0479.880	952.340	1434.044	1916.864	2401.741	2868.722

certain counts, unlike the MMASK which use these counts to mitigate the impact of data disguising process, the Apriori scheme still outperforms the MMASK. However, the MMASK greatly outperforms the MASK since MMASK uses a constant number of counts in order to determine whether an association rule is frequent or not.

To demonstrate the time required for computing association rules within the same data with privacy consideration and different privacy parameters, we run our simulator for different values of w and s_{min}. Particularly, we use the support values from 0.004 to 0.014 with increments of 0.002 and w from 0 to 5 with increments of 1. Table 6 shows the required time of computation in second where shaded measurements represent the time required for commutating association rules when using HS and unshaded measurements represent the time required for computing association rules in PS scheme. The FS scheme's reading are taken when setting $p = 0.5$ and $a = 0.9$.

To compute the memory overhead in each scheme at the same level of resulting privacy, we use w as a weighted memory overhead factor. Results that demonstrate the memory overhead in terms of the parameter w are shown in Table 7. Because it does not use any kind of overhead, $w = 0$ in the PS scheme. For FS scheme, w is computed from Eq. (6) as $w = \frac{1}{1-P_p^{\mathsf{FS}}} - 1$. We also compute w in case of HS scheme according to Eq. (15) as $w = \frac{P_r^{\mathsf{PS}}}{1-P_p^{\mathsf{HS}}} - 1$. For FS, the P_r^{PS} values are realized by setting $p = 0.5$, $p = 0.7$, $p = 0.8$, $p = 0.9$, $p = 0.95$ and $a = 0.9$ [15]. The memory overhead is substituted from Table 7 by multiplying the corresponding w value by the original dataset size. For instance, Table 8 shows the required memory overhead for the dataset T10I8D100k considering different privacy parameters.

Note that the minimal memory in FS is achieved at $P_r^{\mathsf{PS}} = \frac{11}{100}$ which reflects a privacy preservation in PS scheme of 0.89. Because the privacy in PS requires no memory overhead but rather a variation of used parameters a and p, we can always make sure to meet such condition.

Table 7 A comparison between FS, PS, and HS in terms of required w that determines the required memory for each scheme. Note that the values in parenthesis indicate P_r^{PS} used to compute w in P_p^{HS} model which are realized at $p = 0.5$, $p = 0.7$, $p = 0.8$, $p = 0.9$, $p = 0.95$ and $a = 0.9$.

P_p	FS	PS	$(\frac{11}{100})$	$(\frac{12}{100})$	$(\frac{13}{100})$	$(\frac{17}{100})$	$(\frac{23}{100})$
0.91	$10\frac{1}{9}$	0	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{4}{9}$	$\frac{8}{9}$	$1\frac{5}{9}$
0.92	$11\frac{1}{2}$	0	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{5}{8}$	$1\frac{1}{8}$	$1\frac{7}{8}$
0.93	$13\frac{2}{7}$	0	$\frac{4}{7}$	$\frac{5}{7}$	$\frac{6}{7}$	$1\frac{3}{7}$	$2\frac{2}{7}$
0.94	$15\frac{2}{3}$	0	$\frac{5}{6}$	1	$1.\frac{1}{6}$	$1\frac{5}{6}$	$2\frac{5}{6}$
0.95	19	0	$1\frac{1}{5}$	$1\frac{2}{5}$	$1\frac{3}{5}$	$2\frac{2}{5}$	$3\frac{3}{5}$
0.96	24	0	$1\frac{3}{4}$	2	$2\frac{1}{4}$	$3\frac{1}{4}$	$4\frac{3}{4}$
0.97	$32\frac{1}{3}$	0	$2\frac{2}{3}$	3	$3\frac{1}{3}$	$4\frac{2}{3}$	$6\frac{2}{3}$
0.98	49	0	$4\frac{1}{2}$	5	$5\frac{1}{2}$	$7\frac{1}{2}$	$10\frac{1}{2}$
0.99	99	0	10	11	12	16	22

Table 8 Numerical results for the memory requirements in megabytes for storing the dataset T10I8D100k according to the settings in Table 7 for different achievable privacy levels in the different schemes and scenarios.

P_p	FS	PS	$(\frac{11}{100})$	$(\frac{12}{100})$	$(\frac{13}{100})$	$(\frac{17}{100})$	$(\frac{23}{100})$
0.91	44.489	0	0.978	1.4667	1.956	3.911	2.444
0.92	50.600	0	1.650	2.200	2.750	4.950	8.250
0.93	58.457	0	2.514	3.1429	3.771	6.286	10.06
0.94	68.933	0	3.667	4.400	5.133	8.067	12.47
0.95	83.600	0	5.280	6.160	7.040	10.56	15.84
0.96	105.60	0	7.700	8.800	9.900	14.30	20.90
0.97	142.27	0	11.73	13.20	14.67	20.53	29.33
0.98	215.60	0	19.80	22.00	24.20	33.00	46.20
0.99	435.60	0	44.00	48.40	52.80	70.40	96.80

7. Related Works

There have been constant efforts in literature to provide mechanisms that aim at providing PP-ARM in addition to the works discussed in Sect. 3 and Sect. 4. In this section, we touch upon some of these related works.

In [17] a per-transaction noise addition method is introduced for PP-ARM where, given the randomization parameters, a miner can estimate the support and confidence of an association rule over the disguised data. In [16], [33], and [34], selective-hiding methods for 1's and 0's by replacing them with unknowns signs are introduced. Particularly, these works are mostly based on heuristics for estimating the privacy attained by slightly modifying the set of frequent itemsets so that an attacker has less information about *sensitive rules*. In [25], a cryptographic method is introduced for preserving privacy in horizontally partitioned and distributed databases. Similarly, a solution for PP-ARM in a vertically distributed database model is introduced in [14]. In [19], a sanitization method is introduced to sanitize restrictive rules while blocking inference. In [35] a method for mining association rules is introduced in a distributed settings where collusion resistance is provided up to a threshold. A method for blocking anonymity threats raised in frequent itemset mining by limiting inference is introduced in [36]. Finally, in [37], an optimization technique for improving the counting technique in MASK is introduced. Unlike MASK, the introduced optimization technique limits the complexity of counting to 2^c where c is a predefined constant.

8. Conclusion

Privacy preservation in association rule mining (PP-ARM) is a critical issue of research where several works have been proposed for computing the support of itemset in a randomized dataset considering different randomization techniques. In this paper, we revisited the PP-ARM using fake transactions and showed three major results. We first redefined the privacy to include the average case consideration. We then pointed out the exhaustive requirements of the FS in terms of memory and computation. We further pointed out a drawback of the FS in practice by showing its weakness against fake transactions filtering. In order to avoid such limitations of the FS, we extend it to a hybrid scheme with by combining it with the PS scheme and show by both analytical and experimental results the attained properties.

Acknowledgment

We would like also to thank Ku-Young Chang, Jeong Woon Choi, and Ik Rae Jeong for their comments and suggestions that motivated and helped writing this article.

References

[1] K. Wang, P.S. Yu, and S. Chakraborty, “Bottom-up generalization: A data mining solution to privacy protection,” Proc. 1st IEEE International Conference on Data Mining, pp.249–256, 2004.

[2] R. Agrawal and R. Srikant, “Privacy preserving data mining,” Proc. ACM SIGMOD Conference on Management of Data, pp.439–450, Dallas, TX, May 2000.

[3] E. Bertino, I.N. Fovino, and L.P. Provenza, “A framework for evaluating privacy preserving data mining algorithms*,” Data Min. Knowl. Discov., vol.11, no.2, pp.121–154, 2005.

[4] B. Pinkas, “Cryptographic techniques for privacy-preserving data mining,” SIGKDD Explorations, vol.4, no.2, pp.12–19, 2002.

[5] Y. Lindell and B. Pinkas, “Privacy preserving data mining,” Advances in Cryptology (CRYPTO’00), Lect. Notes Comput. Sci., vol.1880, pp.36–53, Springer-Verlag, 2000.

[6] Z. Yang, R. Wright, and H. Subramaniam, “Experimental analysis of a privacy-preserving scalar product protocol,” International Journal of Computer Systems Science & Engineering, vol.21, no.1, pp.47–52, 2006.

[7] Z. Teng and W. Du, “A hybrid multi-group privacy-preserving approach for building decision trees,” Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp.296–307, 2007.

[8] J. Sakuma and S. Kobayashi, “Large-scale k-means clustering with user-centric privacy preservation,” Lect. Notes Comput. Sci., vol.5012, p.320, 2008.

[9] H. Yu, J. Vaidya, and X. Jiang, “Privacy-preserving svm classification on vertically partitioned data,” Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp.647–656, 2006.

[10] S. Guo, X. Wu, and Y. Li, “On the lower bound of reconstruction error for spectral filtering based privacy preserving data mining,” Proc. European Conference on Principles of Data Mining and Knowledge Discovery, pp.520–527, 2006.

[11] S.R.M. Oliveira and O.R. Zaïane, “Achieving privacy preservation when sharing data for clustering,” Proc. Workshop on Secure Data Management, pp.67–82, 2004.

[12] N. Zhang, “Towards comprehensive privacy protection in data clustering,” Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp.1096–1104, 2007.

[13] A. Mohaisen and D. Hong, “Mitigating the ica attack against rotation-based transformation for privacy preserving clustering,” ETRI Journal, vol.30, no.6, pp.1225–6463, 2008.

[14] J. Vaidya and C. Clifton, “Privacy-preserving k-means clustering over vertically partitioned data,” KDD ’03: Proc. Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.206–215, New York, NY, USA, ACM Press, 2003.

[15] S. Rizvi and J.R. Haritsa, “Maintaining data privacy in association rule mining,” Proc. Very Large Data Bases (VLDB) Conference, pp.682–693, 2002.

[16] Y. Saygin, V.S. Verykios, and A.K. Elmagarmid, “Privacy preserving association rule mining,” RIDE, pp.151–158, 2002.

[17] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, “Privacy preserving mining of association rules,” KDD ’02: Proc. Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.217–228, New York, NY, USA, ACM Press, 2002.

[18] J.L. Lin and J.Y.C. Liu, “Privacy preserving itemset mining through fake transactions,” Proc. ACM Symposium on Applied Computing (SAC), pp.375–379, 2007.

[19] S.R.M. Oliveira, O.R. Zaïane, and Y. Saygin, “Secure association rule sharing,” PAKDD, ed. H. Dai, R. Srikant, and C. Zhang, Lect. Notes Comput. Sci., vol.3056, pp.74–85, Springer, 2004.

[20] J.S. Vaidya and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data,” Proc. Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.639–644, Edmonton, Canada, July 2002.

[21] K. Wang, B.C.M. Fung, and P.S. Yu, “Template-based privacy preservation in classification problems,” Proc. 5th IEEE International Conference on Data Mining (ICDM 2005), pp.466–473, Houston, TX, Nov. 2005.

[22] K. Chen and L. Liu, “Privacy preserving data classification with rotation perturbation,” ICDM ’05: Proc. Fifth IEEE International Conference on Data Mining, pp.589–592, Washington, DC, USA, 2005.

[23] D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu, “A fast distributed algorithm for mining association rules,” Proc. Third International Conference on Parallel and Distributed Information Systems, pp.31–42, 1996.

[24] Z. Zheng, R. Kohavi, and L. Mason, “Real world performance of association rule algorithms,” Proc. Knowledge Discovery and Data Mining, pp.401–406, 2001.

[25] M. Kantarcioglu and C. Clifton, “Privacy-preserving distributed mining of association rules on horizontally partitioned data,” Proc. Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD), pp.48–55, 2002.

[26] Department of Health & Human Services, “Health insurance portability and accountability act,” http://www.hhs.gov/ocr/hipaa/, 2009.

[27] PIPEDA, “Privacy commissioner of canada,”
http://www.privcom.gc.ca/, 2009.

[28] European Union Directives, “95/46/EC on the protection of personal data,” www.dataprotection.ie/viewdoc.asp?DocID=92, 2009.

[29] ISO, “International organization for standardization,”
http://www.iso.org/, 2009.

[30] L. Guo, S. Guo, and X. Wu, “Privacy preserving market basket data analysis,” Proc. European Conference on Principles of Data Mining and Knowledge Discovery, pp.103–114, 2007.

[31] R. Agrawal, T. Imielinski, and A.N. Swami, “Mining association rules between sets of items in large databases,” Proc. ACM SIGMOD Conference, pp.207–216, 1993.

[32] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules in large databases,” VLDB’94, Proc. 20th International Conference on Very Large Data Bases, pp.487–499, Santiago de Chile, Chile, Sept. 1994, ed. J.B. Bocca, M. Jarke, and C. Zaniolo, Morgan Kaufmann, 1994.

[33] E. Dasseni, V.S. Verykios, A.K. Elmagarmid, and E. Bertino, "Hiding association rules by using confidence and support," Proc. Information Hiding, 4th International Workshop, pp.369–383, 2001.

[34] V.S. Verykios, A.K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association rule hiding," IEEE Trans. Knowl. Data Eng., vol.16, no.4, pp.434–447, 2004.

[35] J. Wang, T. Fukasawa, S. Urabe, T. Takata, and M. Miyazaki, "Mining frequent patterns securely in distributed system," IEICE Trans. Inf. & Syst., vol.E89-D, no.11, pp.2739–2747, Nov. 2006.

[36] M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi, "Blocking anonymity threats raised by frequent itemset mining," ICDM '05: Proc. Fifth IEEE International Conference on Data Mining, pp.561–564, Washington, DC, USA, IEEE Computer Society, 2005.

[37] P. Andruszkiewicz, "Optimization for mask scheme in privacy preserving data mining for association rules," RSEISP, ed. M. Kryszkiewicz, J.F. Peters, H. Rybinski, and A. Skowron, Lect. Notes Comput. Sci., vol.4585, pp.465–474, Springer, 2007.

Abedelaziz Mohaisen is Ph.D. student at the University of Minnesota Twin Cities. Before that he was a member of Engineering Staff at the Electronics and Telecommunication Research Institute (ETRI) in Korea. He received a B.E. degree in computer engineering from the University of Gaza in 2005 and M.E. degree in information and telecommunication engineering from Inha University, Korea in 2007. His research interests include networks security, data privacy, and cryptography.

Nam-Su Jho received the B.S. degree in mathematics from Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 1999 and the Ph.D. degree in mathematics from Seoul National University, Seoul, Korea, in 2007. Since 2007, he has been with the Electronics and Telecommunications Research Institute as a senior researcher. His research interests include cryptography and information theory.

Dowon Hong received his B.S., M.S. and Ph.D. degrees in mathematics from Korea University, Seoul, Korea on 1994, 1996, and 2000. He is currently a senior member of engineering staff and the team leader of Cryptography Research team at the Electronics and Telecommunication Research Institute, Korea where his research interests are broadly in the area of applied cryptography, networks security, and digital forensics

DaeHun Nyang received the B.Eng. degree in electronic engineering from Korea Advanced Institute of Science and Technology, M.S. and Ph.D. degrees in computer science from Yonsei University, Korea on 1994, 1996, and 2000 respectively. He has been a senior member of engineering staff of Electronics and Telecommunications Research Institute, Korea from 2000 to 2003. Since 2003, he has been an assistant professor of the graduate school of Information Technology and Telecommunication at Inha University, Korea where he is also the founding director of the Information Security Research Laboratory. He is also a consultant for Korean Information Security Agency, member of board of directors and editorial board of Korean Institute of Information Security and Cryptology. Dr. Nyang's research interests include cryptography and information security, privacy, biometrics and their applications to authentication, public key cryptography. Also, he is interested in the security of WLAN, RFID, WSN, and MANET.