Hindawi
Scientific Programming
Volume 2017, Article ID 7545347, 10 pages
https://doi.org/10.1155/2017/7545347

*Research Article*

# Frequent Symptom Sets Identification from Uncertain Medical Data in Differentially Private Way

**Zhe Ding, Zhen Qin, and Zhiguang Qin**

*School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China*

Correspondence should be addressed to Zhen Qin; qinzhen@uestc.edu.cn

Received 16 December 2016; Revised 26 February 2017; Accepted 18 April 2017; Published 11 May 2017

Academic Editor: Tomàs Margalef



Data mining techniques are applied to identify hidden patterns in large amounts of patient data. These patterns can assist physicians in making more accurate diagnosis. For different physical conditions of patients, the same physiological index corresponds to a different symptom association probability for each patient. Data mining technologies based on certain data cannot be directly applied to these patients' data. Patient data are sensitive data. An adversary with sufficient background information can make use of the patterns mined from uncertain medical data to obtain the sensitive information of patients. In this paper, a new algorithm is presented to determine the top $K$ most frequent itemsets from uncertain medical data and to protect data privacy. Based on traditional algorithms for mining frequent itemsets from uncertain data, our algorithm applies sparse vector algorithm and the Laplace mechanism to ensure differential privacy for the top $K$ most frequent itemsets for uncertain medical data and the expected supports of these frequent itemsets. We prove that our algorithm can guarantee differential privacy in theory. Moreover, we carry out experiments with four real-world scenario datasets and two synthetic datasets. The experimental results demonstrate the performance of our algorithm.

## 1. Introduction

The Internet of Things (IoT) involves a lot of different base technologies, such as wireless sensors, data management, and cloud computing [1]. Today, IoT technology is successfully applied in the field of eHealth [2–4]. Medical personnel can utilize IoT technology to collect large amounts of patient data that can assist them in providing better medical services to patients [5, 6].

Frequent itemsets mining is applied in fields such as eHealth and bioinformatics. Traditional algorithms for mining frequent itemsets from medical data are based on certain data [7] and can be applied to discover hidden symptom patterns from a huge amount of data on patient symptoms. These patterns can be used by health managers to provide better healthcare for users [8]. For example, in [9, 10], the Apriori algorithm was applied to identify prevalent diseases and analyze medical billing. However, the Apriori algorithm mines frequent itemsets from certain data. In medicine, for different physical conditions of patients, the same physiological index corresponds to a different symptom association probability for each patient. As a result, there is uncertainty in patient data. Therefore, traditional algorithms for mining frequent itemsets from certain data cannot be directly applied to patient data.

Another important factor is that medical records contain sensitive patient information. An adversary with sufficient background information can make use of frequent patterns mined from patient data to obtain the sensitive information of patients. Hence, it is very important to protect patient privacy when mining frequent itemsets from medical data [11].

The set of symptoms that a patient suffers from constitute the patient's data. Because of the probabilities associated with these symptoms, there is uncertainty in patient data. A large amount of patient data constitutes uncertain data. In the field of medicine, there are plenty of researches on symptom association probability. For example, one study monitored oesophageal pH over a 24 h period to obtain symptom association probability, which was then utilized to evaluate the association between a patient's symptoms and gastroesophageal reflux [12]. By analyzing the large amounts of patient data, Beglinger et al. determined the probability that a patient

suffering from Huntington's disease also had obsessive and compulsive symptoms [13]. By analyzing the data of patients suffering from irritable bowel syndrome, Arsiè et al. determined the probability that indicated the association between meal ingestion and abdominal pain symptoms for patients suffering from irritable bowel syndrome [14]. In this paper, based on symptom association probability obtained by medical technology, we focus on how to mine frequent itemsets from uncertain medical data, while also protecting data privacy. In the uncertain medical data, each item corresponds to a symptom of patients.

In this paper, a new algorithm, denoted as U-PrivMining (uncertain medical data differentially private frequent itemsets mining), is proposed to mine the top $K$ most frequent itemsets from uncertain medical data in a differentially private way. In uncertain medical data, each item corresponds to a symptom of patients. U-PrivMining has two phases. In the first phase, based on traditional algorithms for mining frequent itemsets from uncertain data, spare vector algorithm and the Laplace mechanism are applied to ensure differential privacy for all the frequent itemsets mined from uncertain medical data. In the second phase, based on the frequent itemsets, the Laplace mechanism is applied to ensure differential privacy for the top $K$ most frequent itemsets for uncertain data, as well as the expected supports of these frequent itemsets. We used the spare vector algorithm to improve the efficiency of our algorithm. The spare vector algorithm was used to mine the top $K$ most frequent itemsets from certain data and guaranteed differential privacy in [15]. One major advantage of the spare vector algorithm is that information disclosure affecting differential privacy occurs only for count queries above the threshold; negative answers do not count against the "privacy budget" [15]. The sparse vector algorithm is also suitable for guaranteeing differential privacy when mining frequent itemsets from uncertain data. For certain data, the fixed occurrence counting of an itemset has been applied to determine whether the itemset is frequent. For mining frequent itemsets based on expected support from uncertain data, the expectation of support of an itemset has been utilized to judge whether the itemset is frequent [16]. To summarize, our key contributions are the following:

(i) A new algorithm is proposed to mine the top $K$ most frequent itemsets from uncertain medical data and ensure differential privacy. Traditional algorithms for mining frequent itemsets in differential privacy ways are based on certain data and thus cannot be directly applied to process uncertain medical data.

(ii) Through privacy analysis, we prove that U-PrivMining guarantees differential privacy in theory. Our experimental results on four real-world scenario datasets and two synthetic datasets illustrate the efficiency of U-PrivMining.

This paper is organized as follows. Section 2 presents an overview of related work on eHealth, IoT, frequent itemsets mining for uncertain data, and differential privacy. In Section 3, some notations used in this paper are introduced. The U-PrivMining algorithm and the proof that U-PrivMining satisfies differential privacy in theory are presented in Section 4. In Section 5, the performance of U-PrivMining is evaluated with six datasets. In the last section, we conclude our work.

## 2. Related Works

eHealth applies IoT technology to provide better healthcare services to users. In 2009, Niyato et al. proposed a remote and mobile patient monitoring system that applies heterogeneous wireless access to monitor the biosignals of patient mobility [17]. In 2015, based on the limitations of traditional cellular networks for eHealth services, Yi et al. designed a transmission scheduling mechanism for delay-sensitive medical packets in an eHealth network [18]. The eHealth system based on IoT used monitoring devices to collect large amounts of patient data. Data mining can find hidden patterns in these data, which can assist medical personnel in providing improved medical services to patients. In 2009, Karaolis et al. proposed an algorithm that used mining association rules to assess the risk of coronary events [19]. When traditional data mining technologies are applied to medical data, many useless patterns are discovered. In 2013, Lee et al. proposed a novel algorithm for mining association rule to determine the relationship between blood factors and disease history [20]. This algorithm reduced the number of useless patterns mined from medical data. In 2014, Park et al. used association rules mined from medical data to identify risk behaviors in daily life [21].

The phenomenon of data uncertainty is very common. Traditional algorithms for mining frequent itemsets based on certain data cannot be directly applied to mine frequent itemsets from uncertain data. There are two categories of research on mining frequent itemsets from uncertain data [22]. The first category is mining frequent itemsets based on expected support. In 2007, Chui et al. proposed the notion of expected support and proposed the U-Apriori algorithm based on the Apriori algorithm [23]. The second category is probabilistic frequent itemsets mining. In 2012, the characteristics of Poisson binomial distribution were introduced to mine probabilistic frequent itemsets [24]. In 2012, Bernecker et al. proposed an algorithm based on the frequent pattern tree to mine probabilistic frequent itemsets from uncertain data [25].

Protecting the privacy of patient data is challenge for eHealth and plenty of studies have been conducted on eHealth security [26–33]. Differential privacy can ensure that when one record in the input database of mechanism $A$ is changed, the output of $A$ is insensitive to the change [34]. In 2006, Dwork et al. proposed the Laplace mechanism to ensure differential privacy for real-valued output [35]. In 2010, Bhaskar et al. proposed an algorithm based on truncated frequencies to ensure differential privacy for the top $K$ most frequent itemsets for certain data [36]. In 2012, Li et al. introduced the notion of basis set to ensure differential privacy for mining the top $K$ most frequent itemsets from certain data [37]. In 2014, Lee et al. applied sparse vector algorithm and the Laplace mechanism to guarantee differential privacy for the top $K$ frequent itemsets mined from certain data [15]. In 2015, Su et al. introduced a smart splitting

method to mine frequent itemsets from certain data and ensure differential privacy [38].

Although there are many studies on mining frequent itemsets from certain data in differentially private ways, research on mining frequent itemsets from uncertain data in differentially private ways remains few. This paper focuses on research on mining the top $K$ most frequent itemsets from uncertain data in differentially private ways.

## 3. Preliminaries

The fundamental notions of mining frequent itemsets from uncertain data [23] and differential privacy [34, 35] will be reviewed in this section. These fundamental notions are used throughout this paper. The terms "item" and "symptom" are used interchangeably; "itemset" and "symptom set" can be swapped.

*3.1. Frequent Itemsets Mining for Uncertain Data.* Let $V = \{v_1, v_2, \ldots, v_m\}$ be a set of $m$ items and $T = \{t_1, t_2, \ldots, t_n\}$ as uncertain data with $n$ records. Each record $t_i = \langle (v_{i_1}, P(v_{i_1} \in t_i)), \ldots, (v_{i_k}, P(v_{i_k} \in t_i)) \rangle$ $(1 \le i \le n)$ is a set of uncertain items. For $(v_{i_j}, P(v_{i_j} \in t_i)) \in t_i$, $v_{i_j} \in V$ is assigned with existential probability $P(v_{i_j} \in t_i)$ , which indicates the likelihood that $v_{i_j}$ appears in $t_i$. For example, let $V$ = {hypotension, eating disorder, anemia, neurasthenia}. The uncertain data is shown in Table 1. We can obtain the information from Table 1 as follows. $T = \{t_1, t_2\}$ and $t_1$ = {(hypotension: 0.3), (eating disorder: 0.1)}, which means that user $t_1$ may be suffering from hypotension and eating disorder. The probability of {hypotension} existing in $t_1$ is equal to 0.3; in other words, $P(\text{hypotension} \in t_1) = 0.3$ . This means that the probability of user $t_1$ suffering from hypotension is equal to 0.3.

A set of possible worlds (possible certain database), denoted as $W = \{w_1, w_2, \ldots, w_{|W|}\}$, can be inferred from uncertain data $T$. According to the existing probabilities $P(v_j \in t_i)$, each possible world $w_g$ $(1 \le g \le |W|)$ is illustrated by generating $t_i \in T$. Table 2 shows a set of possible worlds inferred from the uncertain data shown in Table 1. For instance, the possible world $w_2$ = {{hypotension}, {anemia, hypotension}} in Table 2 means that the user $t_1$ is suffering from hypotension and user $t_2$ is suffering from anemia and hypotension.

We assume that all the records in the uncertain data and all the uncertain items in the same record are mutually independent. The probability of a possible world $w_g$, denoted as $P(w_g)$, can be obtained by the following [23]:

$$P(w_g) = \prod_{i=1}^{n} \left( \prod_{x \in T(w_g, t_i)} P(x \in t_i) \cdot \prod_{y \notin T(w_g, t_i)} (1 - P(y \in t_i)) \right), \tag{1}$$

Table 1: Uncertain data.

| ID | Records |
|---|---|
| $t_1$ | (hypotension: 1), (eating disorder: 0.3) |
| $t_2$ | (anemia: 1), (hypotension: 0.7), (neurasthenia: 0.6) |

where $T(w_g, t_i)$ denotes the set of items contained in record $t_i$ and belonging to $w_g$. The expected support of itemset $X$, denoted as $S_e(X)$, can be obtained by the following [23]:

$$S_e(X) = \sum_{i=1}^{|W|} P(w_i) \times S(X, w_i), \tag{2}$$

where $S(X, w_g)$ is the support count of itemset $X$ in possible world $w_g$. For Table 2, in $w_2$, we can obtain the information as $P(w_2) = 1 \times (1-0.3) \times 1 \times 0.7 \times (1-0.6) = 0.196$, $T(w_2, t_1)$ = {hypotension} and $T(w_2, t_2)$ = {anemia, hypotension}.

*3.2. Differential Privacy.* Differential privacy can ensure that output of the analysis mechanism is insensitive to changes in input records. If an analysis mechanism ensures differential privacy, its output will be insensitive to the addition or removal of a record from the input database. As a result, the output cannot be used by adversaries to gain access to a patient's record using their background information [35]. Many studies on privacy protection are based on two assumptions. The first assumption is that the background information of adversaries is already known to the security manager. The second one is that the security manager has known which information should be kept private for users. Differential privacy can protect sensitive information of users without that information [34]. Two databases, $D_1$ and $D_2$, are a pair of neighboring databases if and only if they differ by no more than one record.

*Definition 1* (ϵ-differential privacy [34]). Let Range($A$) be the domain of a random algorithm $A$'s output. $D$ and $D'$ are any pair of neighboring datasets. If (3) is satisfied, then algorithm $A$ guarantees ϵ-differential privacy.

$$P[A(D) = S] \le e^{\epsilon} \cdot P\left[A\left(D'\right) = S\right], \tag{3}$$

where ϵ is the privacy budget of differential privacy and $S \in$ Range($A$).

The sensitivity is used to obtain the maximal possible difference value between outputs for any pair of neighboring datasets.

*Definition 2* (sensitivity [34]). Given the function $f : D^n \to R^d$, the sensitivity of $f$, denoted as $\Delta f$, can be obtained by

$$\Delta f = \max \left\| f(D) - f\left(D'\right) \right\|_1, \tag{4}$$

where $D$ and $D'$ are any pair of neighboring datasets.

*Definition 3* (the Laplace mechanism [35]). Given dataset $D$, let $Q = (q_1, q_2, \ldots, q_p)$ be a query sequence and the sensitivity

Table 2: Possible worlds.

| $W$ | Possible world | $P(w_i)$ |
|---|---|---|
| $w_1$ | {hypotension}, {anemia} | 0.084 |
| $w_2$ | {hypotension}, {anemia, hypotension} | 0.196 |
| $w_3$ | {hypotension}, {anemia, neurasthenia} | 0.126 |
| $w_4$ | {hypotension}, {anemia, hypotension, neurasthenia} | 0.294 |
| $w_5$ | {eating disorder, hypotension}, {anemia} | 0.036 |
| $w_6$ | {eating disorder, hypotension}, {anemia, hypotension} | 0.084 |
| $w_7$ | {eating disorder, hypotension}, {anemia, neurasthenia} | 0.054 |
| $w_8$ | {eating disorder, hypotension}, {anemia, hypotension, neurasthenia} | 0.126 |

of $Q$ is $\Delta Q$. Let $(\xi_1, \xi_2, \ldots, \xi_p)$ be a vector, in which $\xi_i$ $(1 \le i \le p)$ are *i.i.d.* drawn from the Laplace distribution whose scale and mean are $\Delta Q/\epsilon$ and 0, respectively. The algorithm

$$A(D) = Q(D) + \left(\xi_1, \xi_2, \ldots, \xi_p\right) \tag{5}$$

guarantees $\epsilon$-differential privacy.

**Lemma 4** (composition lemma [34]). *Given a sequence of algorithm, denoted as $f = f_1, f_2, \ldots, f_d$, if each algorithm $f_i$ $(1 \le i \le d)$ guarantees $\epsilon_i$-differential privacy, then $f$ ensures $\sum_{i=1}^{d} \epsilon_i$-differential privacy.*

## 4. U-PrivMining Algorithm

This section introduces the U-PrivMining algorithm to determine the top $K$ most frequent itemsets from uncertain data, in which each item corresponds to a symptom of patients, in a differentially private way. The process of U-PrivMining consists of two phases. In the first phase, the assigned privacy budget is equal to $\epsilon_1 = \alpha \cdot \epsilon$. In the second phase, the assigned privacy budget is equal to $\epsilon_2 = (1 - \alpha) \cdot \epsilon$. The parameter $\alpha \in (0, 1)$ is applied to control the value of the privacy budgets assigned in the two phases. In this study, we chose $\alpha = 1/3$ for all uncertain data. However, this choice may not be optimal. It appears that the optimal allocation depends on the characteristics of the uncertain medical data and value of $K$ [36].

*4.1. Description of U-PrivMining.* The whole process of U-PrivMining is introduced in this section. U-PrivMining is composed of two phases. In the first phase, we can obtain $\theta$ so that the expected supports of the top $K$ most frequent itemsets are greater than or equal to $\theta$. The privacy budget allocated to this step is equal to $\epsilon_1 = (1/3) \cdot \epsilon$. On the basis of traditional algorithm for mining frequent itemsets from uncertain data, we apply the sparse vector algorithm [15] and Laplace mechanism to ensure $(\epsilon/3)$-differential privacy for this phase. The steps in the first phase of U-PrivMining are as follows.

*Step 1.* The expected support of the $K$th most frequent itemset, denoted as $S_k$, is obtained by utilizing traditional algorithms for mining frequent itemsets based on expected support from uncertain data.

*Step 2.* The noisy threshold, denoted as $\widehat{S}_K$, can be obtained by

$$\widehat{S}_K = S_K + \mathrm{Lap}\left(\frac{12}{\epsilon}\right), \tag{6}$$

where $\mathrm{Lap}(12/\epsilon)$ is the noisy data generated by the Laplace distribution, whose mean and scale are 0 and $(12/\epsilon)$, respectively.

*Step 3.* On the basis of traditional algorithms for mining frequent itemsets from uncertain data, the sparse vector algorithm is applied to obtain all the frequent itemsets whose assessment expected supports are greater than or equal to the noisy threshold $\widehat{S}_K$. The assessment expected support of an itemset $X$, denoted as $\widehat{S}_e(X)$ can be obtained by

$$\widehat{S}_e(X) = S_e(X) + \mathrm{Lap}\left(\frac{4}{\epsilon}\right), \tag{7}$$

where $S_e(X)$ is the expected support of itemset $X$ and $\mathrm{Lap}(4/\epsilon)$ is the noisy data generated by the Laplace distribution, whose mean and scale are 0 and $(4/\epsilon)$, respectively.

*Step 4.* All the frequent itemsets obtained in Step 3 and the expected supports of these itemsets are taken as the output of this phase.

In the second phase, according to the output of the first phase, U-PrivMining can obtain the top $K$ most frequent itemsets for uncertain data and the noisy expected supports of these frequent itemsets. The privacy budget allocated to the second phase is equal to $\epsilon_2 = (2/3) \cdot \epsilon$. The privacy budgets allocated to ensure differential privacy for the top $K$ most frequent itemsets for uncertain data and for the expected supports of these itemsets for uncertain data are equal to $\epsilon_{2,1} = \beta \cdot \epsilon_2$ and $\epsilon_{2,2} = (1 - \beta) \cdot \epsilon_2$, respectively. The second phase of U-PrivMining is described below.

Let $H = \{h_1, h_2, \ldots, h_{|H|}\}$ be a set of itemsets obtained in the first phase of U-PrivMining. Let $S_e(h_i)$ $(1 \le i \le |H|)$ be the expected support of itemset $h_i$. The steps in the second phase of U-PrivMining are as follows.

*Step 1* (if $|H|$ is less than or equal to $K$, $\beta$ is equal to 0). All the itemsets in $H$ belong to the top $K$ most frequent itemsets for uncertain data. And then Step 3 is directly executed.

*Step 2* (if $|H|$ is greater than $K$, $\beta$ is equal to 0.5). The perturbation expected supports of all the itemsets in $H$ can be obtained. The perturbation expected support of itemset $h_i$ $(1 \le i \le |H|)$, denoted as $\varsigma(H_i)$, can be obtained by

$$\varsigma(h_i) = S_e(h_i) + \xi_i, \tag{8}$$

where $\xi_1, \xi_2, \ldots, \xi_{|H|}$ are mutually independent and drawn from the Laplace distribution, whose mean and scale are 0 and $(|H|/\epsilon_{2,1})$, respectively. The top $K$ most frequent itemsets for the perturbation expected supports in $H$ are the top $K$ most frequent itemsets for uncertain data.

*Step 3.* Let $R = \{r_1, r_2, \ldots, r_K\}$ be the set of the top $K$ most frequent itemsets for uncertain data, which are obtained in above steps. The noisy expected supports of all the itemsets in $R$ can be obtained. The noisy expected support of itemset $r_i$ $(1 \le i \le K)$ can be obtained by

$$\sigma(r_i) = S_e(r_i) + \lambda_i, \tag{9}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_K$ are mutually independent and drawn from the Laplace distribution whose mean and scale are equal to 0 and $(K/\epsilon_{2,2})$, respectively.

*Step 4.* The top $K$ most frequent itemsets for uncertain data and the noisy expected supports of these itemsets are taken as the output of U-PrivMining.

*4.2. Privacy Analysis for U-PrivMining.* In this section, we prove that U-PrivMining is $\epsilon$-differentially private. In order to prove that U-PrivMining guarantees differential privacy, we introduce the notions of count query set and threshold query set.

*Definition 5* (count query set [15]). Let $S = \{s_1, s_2, \ldots, s_{|S|}\}$ be a set of itemsets with $|S|$ itemsets. A count query set is composed of a number of queries. Let $CQ = (q_1, q_2, \ldots, q_{|S|})$ be the count query set, where each query $q_i$ $(1 \le i \le |S|)$ asks for the expected support of the $i$th itemset in $S$.

*Definition 6* (threshold query set [15]). Let $S = \{s_1, s_2, \ldots, s_{|S|}\}$ be a set of itemsets with $|S|$ itemsets. A threshold query set is composed of a number of threshold queries. Let $TQ = (q_1, q_2, \ldots, q_{|S|})$ be the threshold query set, where each $q_i$ $(1 \le i \le |S|)$ returns 1 if $\widehat{S}(s_i) \ge \widehat{S}_K$ $(1 \le i \le |S|)$; otherwise $q_i$ returns 0.

According to the definition of count query set, the sensitivity of the count query and count query set can be obtained as follows.

**Lemma 7.** *Let $CQ = (q_1, q_2, \ldots, q_{|H|})$ be a count query set. The sensitivity of $q_i$ $(1 \le i \le |H|)$ and $CQ$ are equal to 1 and $|H|$, respectively.*

*Proof.* According to (2), we can obtain the other method to compute the expected support of an itemset $X$, denoted as $S_e(X)$, as follows [23]:

$$S_e(X) = \sum_{i=1}^{n} \prod_{x \in X} P(x \in t_i), \tag{10}$$

where $n$ is the number of records in an uncertain data $T$ and $t_i$ $(1 \le i \le n)$ is a record in $T$. Let $T$ and $T'$ be a pair of neighbor databases. Let $D = \{t \mid t \in T \cap t \in T'\}$ be the intersection of $T$ and $T'$. Let $|T|$, $|T'|$, and $|D|$ be the total size of $T$, $T'$, and $D$, respectively. Let $S_e^T(X)$ and $S_e^{T'}(X)$ be the expected supports of itemset $X$ for $T$ and $T'$, respectively. According to (10), the values of $S_e^T(X)$ and $S_e^{T'}(X)$ can be computed as follows:

$$\begin{aligned} S_e^T(X) &= \sum_{j=1}^{|T|} \prod_{x \in X} P(x \in t_j) \\ &= \sum_{j=1}^{|D|} \prod_{x \in X} P(x \in t_j) + \prod_{x \in X} P(x \in d_1) \\ S_e^{T'}(X) &= \sum_{j=1}^{|T'|} \prod_{x \in X} P(x \in t_j) \\ &= \sum_{j=1}^{|D|} \prod_{x \in X} P(x \in t_j) + \prod_{x \in X} P(x \in d_2) \\ &\Downarrow \\ \left\| S_e^T(X) - S_e^{T'}(X) \right\|_1 &\le 1, \end{aligned} \tag{11}$$

where $d_1 = \{t \mid t \in T \cap t \notin T'\}$ and $d_2 = \{t \mid t \in T' \cap t \notin T\}$. As a result, the sensitivity of each query $q_i$ $(1 \le i \le |H|)$ is equal to 1. Since there are $|H|$ queries in $CQ$, the sensitivity of $CQ$ is equal to $|H|$. □

Based on the sensitivity of the count query and count query set for uncertain data, we can conclude that U-PrivMining guarantees $\epsilon$-differential privacy. The proof procedure is outlined below.

**Theorem 8.** *The first phase of U-PrivMining is $(\epsilon/3)$-differentially private.*

*Proof.* According to Lemma 7, the sensitivity of obtaining $S_K$ is equal to 1. As a result, according to the Laplace mechanism, it is $(\epsilon/12)$-differentially private to generate the noisy threshold $\widehat{S}_K$. Let $\widehat{S}_K^{D_1}$ and $\widehat{S}_K^{D_2}$ be the noisy threshold for a pair of neighboring databases $D_1$ and $D_2$, respectively. According to Definition 1, (12) is satisfied.

$$P\left(\widehat{S}_K^{D_1} = x\right) \le P\left(\widehat{S}_K^{D_2} = x\right) \cdot e^{(\epsilon/12)}. \tag{12}$$

Let $H = \{h_1, h_2, \ldots, h_{|H|}\}$ be the set of itemsets. The threshold query set is applied to model the set of answers as a vector $Q = (q_1, q_2, \ldots, q_{|H|})$ where $q_i = 1$ $(1 \le i \le |H|)$ if $\widehat{S}_e(h_i) \ge \widehat{S}_K$ $(1 \le i \le |H|)$; otherwise $q_i = 0$. Given any pair of neighboring databases $D_1$ and $D_2$, $V_1$ and $V_2$ denote the

output distribution on $Q$ when $D_1$ and $D_2$ are input neighbor databases, respectively. Then, (13) is satisfied (the details of the proof are shown in [15]).

$$\frac{V_1(Q)}{V_2(Q)} \leq e^{(\epsilon/3)}. \tag{13}$$

Thus, the first phase of U-PrivMining ensures $(\epsilon/3)$-differential privacy. □

**Theorem 9.** *The second phase of U-PrivMining is* $(2\epsilon/3)$*-differentially private.*

*Proof.* According to Lemma 7, the sensitivity of obtaining the expected support of an itemset is equal to 1. Therefore, the sensitivity of obtaining the expected support of all frequent itemsets in $H$ is equal to $|H|$. In $|H|$ that is greater than $K$, according to the Laplace mechanism, the scale of the Laplace distribution, which is used to ensure differential privacy for the top $K$ most frequent itemsets, is equal to $(|H|/\epsilon_{2,1})$. Hence, obtaining the top $K$ most frequent itemsets ensures $\epsilon_{2,1}$-differential privacy. The sensitivity of obtaining the expected supports of the top $K$ most frequent itemsets is equal to $K$. According to the Laplace mechanism, the noisy data, which is used to obtain the noisy expected support of the top $K$ frequent itemsets, obeys the Laplace distribution whose scale is equal to $(K/\epsilon_{2,2})$. Hence, it ensures $\epsilon_{2,2}$-differential privacy for obtaining noisy expected supports of the top $K$ most frequent itemsets for uncertain data. As a consequence, according to Lemma 4, the second phase of U-PrivMining guarantees $(2\epsilon/3)$-differential privacy. □

According to analysis of the two phases of U-PrivMining, we can conclude that the first and second phases are $(\epsilon/3)$-differentially private and $(2\epsilon/3)$-differentially private, respectively. According to Lemma 4, U-PrivMining is $\epsilon$-differentially private.

## 5. Experiments

In our experiments, four real-world scenario datasets and two synthetic datasets were utilized to verify the efficiency of U-PrivMining, which can be downloaded from [39]. The parameters of these public datasets are shown in Table 3, where the number of items in the datasets is denoted as $m$ and the number of transactions in the dataset is denoted as $n$. The maximal length of transactions in the dataset is denoted as $\max|t|$. The average length of transactions in the dataset is denoted as $\text{avg}|t|$. In order to add uncertainty to these datasets, an existential random probability in the range of $[0, 1]$ is assigned to each item in each transaction.

Table 3: Dataset.

| Dataset | $m$ | $n$ | $\max\|t\|$ | $\text{avg}\|t\|$ |
|---|---|---|---|---|
| Accidents | 468 | 340183 | 51 | 33.8 |
| Kosarak | 41270 | 990002 | 2498 | 8.1 |
| Pumsb | 2113 | 49046 | 74 | 74 |
| Pumsb star | 2088 | 49046 | 63 | 50.5 |
| T10I4D100K | 870 | 100000 | 29 | 10.1 |
| T40I10D100K | 942 | 100000 | 77 | 39.6 |

*5.1. Evaluation Metrics.* U-PrivMining applies the Laplace mechanism and the spare vector algorithm to ensure differential privacy for the top $K$ most frequent itemsets for uncertain data and the expected supports of these frequent itemsets. The Laplace mechanism can protect the privacy of U-PrivMining's output by adding noisy data to the output of mining frequent itemsets from uncertain data. Thus, the $F$-score and relative error (RE) are applied to evaluate the influence of noisy data on the experimental results.

*Definition 10* ($F$-score [15]). Let $F$ be the set of the top $K$ most frequent itemsets for uncertain data and $\widehat{F}$ be the set of the frequent itemsets obtained by U-PrivMining. The $F$-score can be obtained by

$$F\text{-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{14}$$

where $\text{precision} = |F \cap \widehat{F}|/|\widehat{F}|$ and $\text{recall} = |F \cap \widehat{F}|/|F|$.

*Definition 11* (relative error [15]). Let $S_e(X)$ and $\sigma(X)$ be the expected support of itemset $X$ for uncertain data and the noisy expected support of itemset $X$, respectively, which is obtained in the second phase of U-PrivMining. The RE can be obtained by

$$\text{RE} = \text{median}_{X \in \widehat{F}} \frac{\left|S_e(X) - \sigma(X)\right|}{S_e(X)}. \tag{15}$$

As described in Definition 10, for all the itemsets mined by U-PrivMining, the precision is utilized to evaluate the proportion of itemsets mined by U-PrivMining and belonging to the correct top $K$ most frequent itemsets for uncertain data. The recall is also used to evaluate the proportion of itemsets mined by U-PrivMining and belonging to the correct top $K$ most frequent itemsets for uncertain data. The $F$-score is the harmonic mean of both precision and recall. When the number of the frequent itemsets obtained from the first phase of U-PrivMining is greater than or equal to $K$, the value of $|F|$ and $|\widehat{F}|$ is equal to $K$. As a result, the value of $F$-score and recall is equal to the value of precision.

As described in Definition 11, the value of RE is utilized to evaluate the influence of the noisy data on the noisy expected supports of the top $K$ most frequent itemsets for uncertain data. There may be extremely large or small values in the experimental results. The median was not skewed because these values were extremely large or small. Therefore, the median was applied to evaluate the relative error.

*5.2. Analysis of Experimental Results.* U-PrivMining can identify the top $K$ most frequent itemsets from uncertain data in differentially private way. In traditional algorithms for mining the top $K$ most frequent itemsets from uncertain data and certain data, the $K$ values were predetermined by users

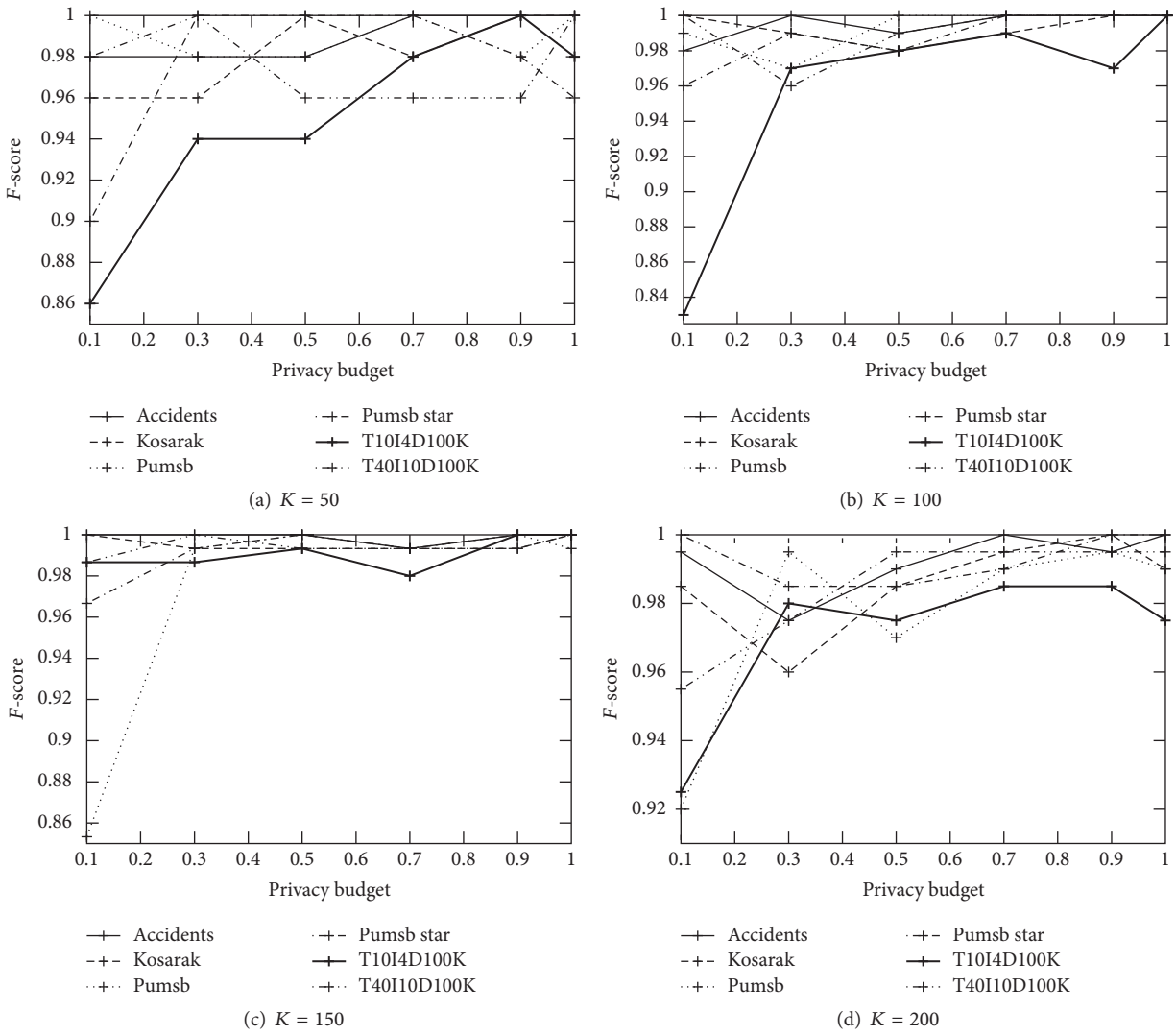


Figure 1: $F$-score by varying privacy budget.

or domain experts [7]. In order to evaluate the influence of privacy budget on the $F$-score and RE, we conducted four group experiments. The $K$ values were set as 50, 100, 150, and 200, respectively.

Figure 1 shows the results of the $F$-score obtained by U-PrivMining running on the six public datasets under different privacy budget values. As it can be seen from the figure, when $K$ value is fixed, the $F$-score fluctuates and is close to 1 with increasing privacy budget. In the first phase of U-PrivMining, the algorithm obtains noisy data to generate noisy threshold and assessment expected supports of itemsets. According to (6) and (7), the greater the value of the privacy budget, the smaller the scale of Laplace distribution used to generate the noisy data in this step. In the second phase of U-PrivMining, the algorithm can obtain the top $K$ most frequent itemsets by adding noisy data to the expected support. The noisy data is drawn from the Laplace distributions, whose mean and scale are equal to 0 and $(H/\epsilon_{2,1})$, respectively. As a result, the $F$-score improves and is close to 1 with increasing privacy budget. From Figure 1, we can conclude that the lower the expected supports of the top $K$ most frequent itemsets for the uncertain data, the lower the convergence speed of the $F$-score. For the T10I4D100K dataset, the expected supports of the top $K$ most frequent itemsets are less than other datasets. Therefore, the convergence speed of U-PrivMining running on the T10I4D100K data set is lower than that of U-PrivMining running on the other datasets. U-PrivMining applied the Laplace mechanism to ensure data privacy. Hence, if the noisy data is relatively greater for the expected supports of the top $K$ most frequent itemsets, then the $F$-score of U-PrivMining is relatively lower.

Figure 2 shows the RE results obtained by U-PrivMining running on six public datasets under different privacy budget values. When $K$ is a fixed value, with increasing privacy budget, the value of RE fluctuates and is close to 0. The noisy expected support of an itemset is obtained by adding the noisy data drawn from the Laplace distribution to the expected support of the itemset. As a consequence, when the

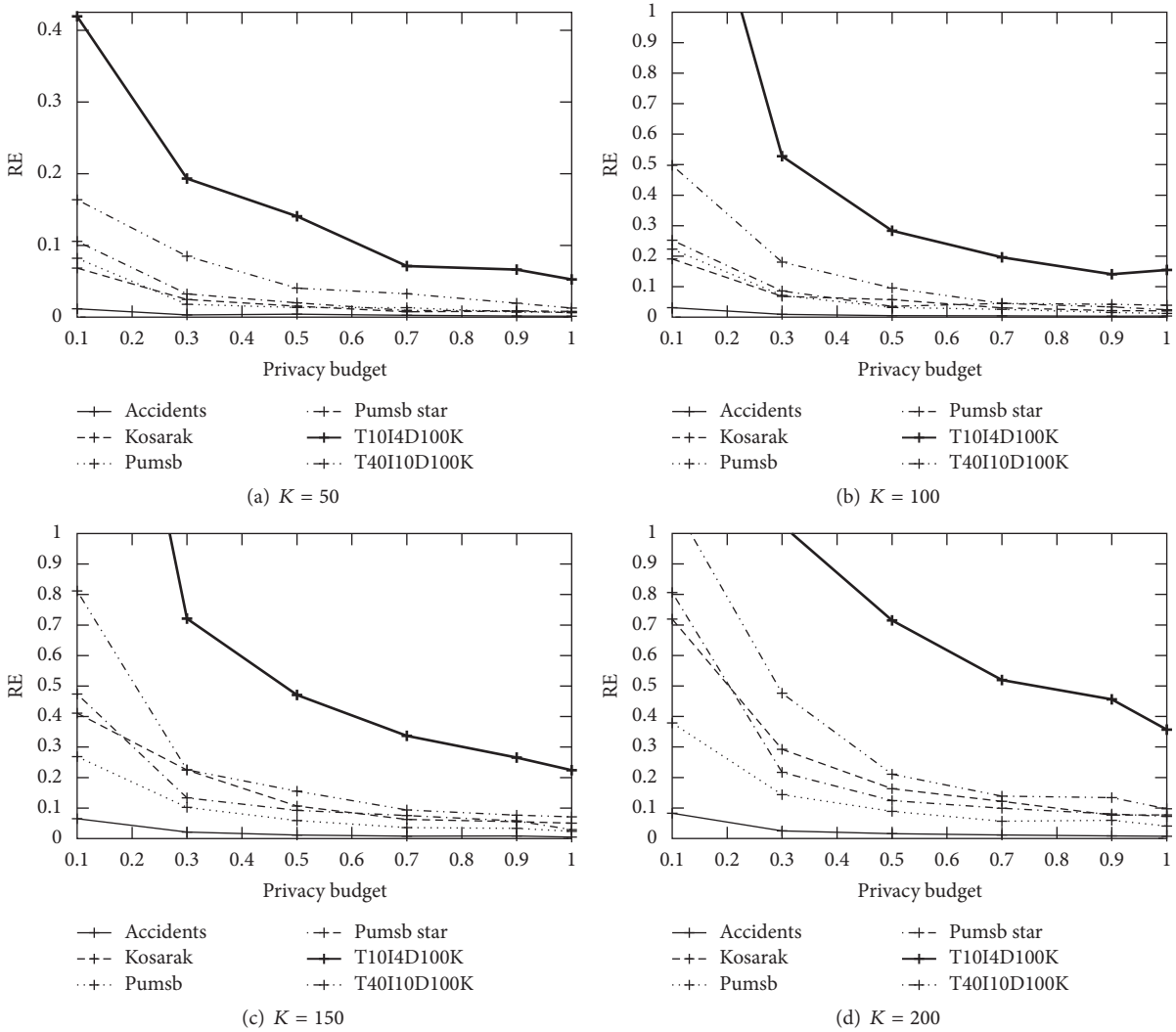


Figure 2: RE by varying privacy budget.

$K$ and privacy budget values are fixed values, the expected supports of the top $K$ most frequent itemsets for the different datasets are lower, and the RE of U-PrivMining is higher. When the privacy budget is a fixed value, and, with increasing $K$, the lower expected supports of the top $K$ most frequent itemsets for different datasets, the higher RE of U-PrivMining. For the same dataset and privacy budget, RE values increase with increasing $K$. The noisy expected support of an itemset can be obtained by adding the noisy data drawn from the Laplace distribution to the expected support of the itemset.

*5.3. Discussion.* In the field of medicine, for different physical conditions of patients, the same physiological index corresponds to a different symptom association probability for each patient. There are plenty of medical technologies to obtain symptom association probability for patients. There is uncertainty in patient data. However, existing algorithms for mining frequent itemsets from medical data in differentially private ways are all based on certain data and cannot be directly used for uncertain medical data. Therefore, in this paper, we proposed the U-PrivMining algorithm, which can mine the top $K$ most frequent itemsets from uncertain medical data and ensure differential privacy. The experimental results verified the effectiveness of U-PrivMining.

## 6. Conclusion

In this paper, we proposed a new algorithm to mine the top $K$ most frequent itemsets from uncertain medical data, where each item corresponds to a patient symptom, while protecting data privacy. These frequent itemsets can assist physicians in making diagnoses. Through theoretical and experimental analyses, we can conclude that not only does U-PrivMining ensure differential privacy but, with increasing privacy budget, the top $K$ most frequent itemsets obtained by U-PrivMining and the noisy expected supports of these frequent itemsets are close to the true top $K$ most frequent itemsets and expected supports of these itemsets for uncertain data, respectively. However, the privacy budget allocation may not

be optimal. The optimization of privacy budget allocation will be focus of future research.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

This work was supported in part by the National Science Foundation of China (no. 61672135, no. 61502085, no. 61272527, and no. 61370026), the National High Technology Research and Development Program of China (no. 2015AA016007), China Postdoctoral Science Foundation Funded Project (no. 2015M570775), the Sichuan Science-Technology Support Plan Program (no. 2014GZ0106, no. 2015GZ0095, and no. 2016JZ0020), and the National Science Foundation of China-Guangdong Joint Foundation (no. U1401257).

## References

[1] J. P. Conti, "The internet of things," *IET Communications Engineer*, vol. 4, no. 6, pp. 20–25, 2006.

[2] H. Abie and I. Balasingham, "Risk-based adaptive security for smart iot in ehealth," in *Proceedings of the 7th International Conference on Body Area Networks*, pp. 269–275, Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, Oslo, Norway, 2012.

[3] H. Oh, C. Rizo, M. Enkin et al., "What is eHealth (3): a systematic review of published definitions," *Journal of Medical Internet Research*, vol. 7, no. 1, 2005.

[4] L. Atzori, A. Iera, and G. Morabito, "The internet of things: a survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[5] A. S. Koyuncugil and N. Ozgulbas, "Donor research and matching system based on data mining in organ transplantation," *Journal of Medical Systems*, vol. 34, no. 3, pp. 251–259, 2010.

[6] C. H. McCollough, "Automated data mining of exposure information for dose management and patient safety initiatives in medical imaging," *Radiology*, vol. 264, no. 2, pp. 322–324, 2012.

[7] J. Han, J. Pei, and M. Kamber, *Data Mining Concepts and Techniques*, Elsevier, 2012.

[8] N. Jay, F. Kohler, and A. Napoli, "Using formal concept analysis for mining and interpreting patient flows within a healthcare network," in *Concept Lattices and Their Applications*, pp. 263–268, Springer, Berlin, Heidelberg, 2008.

[9] M. Ilayaraja and T. Meyyappan, "Mining medical data to identify frequent diseases using Apriori algorithm," in *Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on. IEEE*, pp. 194–199, 2013.

[10] U. Abdullah, J. Ahmad, and A. Ahmed, "Analysis of effectiveness of apriori algorithm in medical billing data mining," in *Emerging Technologies, 2008. ICET 2008. 4th International Conference*, pp. 327–331, 2008.

[11] B. Milovic, "Prediction and decision making in Health Care using Data Mining," *International Journal of Public Health Science*, vol. 1, no. 2, 2012.

[12] S. A. Taghavi, M. Ghasedi, M. Saberi-Firoozi et al., "Symptom association probability and symptom sensitivity index: preferable but still suboptimal predictors of response to high dose omeprazole," *Gut*, vol. 54, no. 8, pp. 1067–1071, 2005.

[13] L. J. Beglinger, D. R. Langbehn, K. Duff et al., "Probability of obsessive and compulsive symptoms in Huntington's disease," *Biological Psychiatry*, vol. 61, no. 3, pp. 415–418, 2007.

[14] E. Arsiè, M. Coletta, B. M. Cesana et al., "Symptom-association probability between meal ingestion and abdominal pain in patients with irritable bowel syndrome. Does somatization play a role?" *Neurogastroenterology and Motility*, vol. 27, no. 3, pp. 416–422, 2015.

[15] J. Lee and C. W. Clifton, "Top-$k$ frequent itemsets via differentially private FP-tree," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*, pp. 931–940, New York, NY, USA, August 2014.

[16] Y. Tong, L. Chen, Y. Cheng et al., "Mining frequent itemsets over uncertain databases," *Proceedings of the Vldb Endowment*, vol. 5, no. 11, pp. 1650–1661, 2012.

[17] D. Niyato, E. Hossain, and S. Camorlinga, "Remote patient monitoring service using heterogeneous wireless access networks: architecture and optimization," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 4, pp. 412–423, 2009.

[18] C. Yi, A. S. Alfa, and J. Cai, "An incentive-compatible mechanism for transmission scheduling of delay-sensitive medical packets in e-health networks," *IEEE Transactions on Mobile Computing*, vol. 15, no. 10, pp. 2424–2436, 2016.

[19] M. Karaolis, J. A. Moutiris, L. Papaconstantinou, and C. S. Pattichis, "Association rule analysis for the assessment of the risk of coronary heart events.," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'09)*, pp. 6238–6241, Conference PubMed, 2009.

[20] D. G. Lee, K. S. Ryu, M. Bashir, J. W. Bae, and K. H. Ryu, "Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction," *Journal of Medical Systems*, vol. 37, no. 2, pp. 1–10, 2013.

[21] S. H. Park, S. Y. Jang, H. Kim, and S. W. Lee, "An association rule mining-based framework for understanding lifestyle risk behaviors," *PLoS ONE*, vol. 9, no. 2, Article ID e88859, 2014.

[22] Y. Tong, L. Chen, Y. Cheng, and P. S. Yu, "Mining frequent itemsets over uncertain databases," *Proceedings of the VLDB Endowment*, vol. 5, no. 11, pp. 1650–1661, 2012.

[23] C. K. Chui, B. Kao, and E. Hung, "Mining frequent itemsets from uncertain data," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 47–58, Springer, Berlin Heidelberg, 2007.

[24] L. Wang, D. W. L. Cheung, R. Cheng, S. D. Lee, and X. S. Yang, "Efficient mining of frequent item sets on large uncertain databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 12, pp. 2170–2183, 2012.

[25] T. Bernecker, H. P. Kriegel, M. Renz, F. Verhein, and A. Züfle, "Probabilistic frequent pattern growth for itemset mining in uncertain databases," in *International Conference on Scientific and Statistical Database Management*, Springer, Berlin, Heidelberg, 2012.

[26] D. Chen, Z. Qin, X. Mao, P. Yang, Z. Qin, and R. Wang, "SmokeGrenade: an efficient key generation protocol with artificial interference," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 11, pp. 1731–1745, 2013.

[27] K. Zhang, X. Liang, R. Lu, and X. Shen, "Sybil attacks and their defenses in the internet of things," *IEEE Internet of Things Journal*, vol. 1, no. 5, pp. 372–383, 2014.

[28] N. Zhang, N. Cheng, N. Lu, X. Zhang, J. W. Mark, and X. Shen, "Partner selection and incentive mechanism for physical layer security," *IEEE Transactions on Wireless Communications*, vol. 14, no. 8, pp. 4265–4276, 2015.

[29] D. Chen, N. Zhang, Z. Qin et al., "S2M: a lightweight acoustic fingerprints based wireless device authentication protocol," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 88–100, 2017.

[30] K. Yang, Q. Han, H. Li, K. Zheng, Z. Su, and X. Shen, "An efficient and fine-grained big data access control scheme with privacy-preserving policy," *IEEE Internet of Things Journal*, 1 page, 2016.

[31] N. Zhang, N. Lu, N. Cheng et al., "Cooperative spectrum access towards secure information transfer for CRNs," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 11, pp. 2453–2464, 2013.

[32] D. Chen, S. Jiang, and Z. Qin, "Message Authentication Code over a wiretap channel," in *IEEE International Symposium on Information Theory*, pp. 2301–2305, 2015.

[33] Q. Wang, D. Chen, N. Zhang et al., "LACS: A Lightweight Label-Based Access Control Scheme in IoT-Based 5G Caching Context," *IEEE Access*, 4027 pages, 2017.

[34] C. Dwork, "The differential privacy frontier," in *Proceedings of Theory of Cryptography Conference*, pp. 496–502, Springer, Berlin Heidelberg, 2009.

[35] C. Dwork, F. McSherry, and K. Nissim, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, pp. 265–284, Springer, Berlin Heidelberg, 2006.

[36] R. Bhaskar, S. Laxman, A. Smith et al., "Discovering frequent patterns in sensitive data," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 503–512, 2010.

[37] N. Li, W. Qardaji, D. Su et al., "Privbasis: frequent itemset mining with differential privacy," *Proceedings of the Vldb Endowment*, vol. 5, no. 11, pp. 1340–1351, 2012.

[38] S. Su, S. Xu, X. Cheng et al., "Differentially private frequent itemset mining via transaction splitting," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 7, pp. 1875–1891, 2015.

[39] "Frequent itemset mining dataset repository," http://fimi.ua.ac.be/data/.