Mining Probabilistic Representative Frequent Patterns From Uncertain Data

Chunyang Liu∗, Ling Chen†, Chengqi Zhang‡

Abstract
Probabilistic frequent pattern mining over uncertain data has received a great deal of attention recently due to the wide applications of uncertain data. Similar to its counterpart in deterministic databases, however, probabilistic frequent pattern mining suffers from the same problem of generating an exponential number of result patterns. The large number of discovered patterns hinders further evaluation and analysis, and calls for the need to find a small number of representative patterns to approximate all other patterns. This paper formally defines the problem of probabilistic representative frequent pattern (P-RFP) mining, which aims to find the minimal set of patterns with sufficiently high probability to represent all other patterns. The problem's bottleneck turns out to be checking whether a pattern can probabilistically represent another, which involves the computation of a joint probability of supports of two patterns. To address the problem, we propose a novel and efficient dynamic programming-based approach. Moreover, we have devised a set of effective optimization strategies to further improve the computation efficiency. Our experimental results demonstrate that the proposed P-RFP mining effectively reduces the size of probabilistic frequent patterns. Our proposed approach not only discovers the set of P-RFPs efficiently, but also restores the frequency probability information of patterns with an error guarantee.

1 Introduction
Uncertainty is inherent in data from many different domains, including sensor network monitoring, moving object tracking, and protein-protein interaction data [6]. Instead of cleaning uncertain data by considering only the most possible circumstance, it is more reasonable to model the uncertainty of data. Consequently, data mining over uncertain data has become an active area of research in recent years. A survey of state-of-the-art uncertain data mining techniques can be found in [1]. As one of the most fundamental data mining tasks, frequent pattern mining over uncertain data has also received a great deal of research attention, since it was first introduced in [3]. Currently, there exist two different definitions of frequent patterns in the context of uncertain data: expected support-based frequent patterns [3, 11], and probabilistic frequent patterns [4, 5]. Both definitions consider the support of a pattern as a discrete random variable. The former uses the expectation of the support as the measurement, while the latter considers the probability that the support of a pattern is no less than some specified minimum support threshold. Various algorithms have been designed to mine frequent patterns from uncertain data. A summarization and comparison of eight algorithms, proposed to mine respectively the two types of aforementioned frequent patterns from uncertain databases, have been reported recently in [6].

Note that, the anti-monotonic property holds for both the expected support-based frequent patterns, as well as the probabilistic frequent patterns. That is, if a pattern is frequent in an uncertain database, then all of its sub-patterns are frequent as well. This property leads to the generation of an exponential number of result patterns. The large number of discovered frequent patterns makes the understanding of, and further analysis of generated patterns troublesome. Therefore, it is important to find a small number of representative patterns to best approximate all other patterns.

Some initial research work has been undertaken to find a small set of representative patterns. For example, mining probabilistic frequent closed patterns over uncertain data has been studied in [7, 8, 9]. However, the number of probabilistic frequent closed patterns is still large because of the restrictive condition for a pattern being closed. For instance, in [9], the closed probability of a pattern is computed as the sum of the probabilities of the possible worlds of an uncertain database where the pattern is closed. In this work, we aim to relax the restrictive condition to further reduce
the size of frequent patterns mined over uncertain data.

In the context of deterministic data, a pattern is closed if it is the longest pattern that appears in the same set of transactions supporting its sub-patterns. As a generalization of the concept of frequent closed patterns, Xin et al. [20] proposed the notion of a \( \varepsilon \)-covered relationship between patterns. A pattern \( X_1 \) is \( \varepsilon \)-covered by another pattern \( X_2 \) if \( X_1 \) is a subset of \( X_2 \) and \( (\text{supp}(X_1) - \text{supp}(X_2))/\text{supp}(X_1) \leq \varepsilon \). The goal is then to find a minimum set of representative patterns that can \( \varepsilon \)-cover all frequent patterns. Since the support of a pattern \( (\text{supp}(X)) \) becomes a discrete random variable in an uncertain database, the \( \varepsilon \)-covered relationship cannot be applied directly to probabilistic frequent patterns.

In this work, we first extend the concept of \( \varepsilon \)-covered by defining a new \((\varepsilon,\delta)\)-covered relationship between probabilistic frequent patterns. Informally, a pattern \( X_1 \) is \((\varepsilon,\delta)\)-covered by another pattern \( X_2 \) in an uncertain database if \( X_1 \) is a subset of \( X_2 \), and the probability that the support distance between \( X_1 \) and \( X_2 \) is no greater than \( \varepsilon \) is no less than \( \delta \). The objective of probabilistic representative frequent pattern \((P-RFP)\) mining is then to find the minimal set of patterns that can \((\varepsilon,\delta)\)-cover all probabilistic frequent patterns.

The approach for P-RFPs mining can be divided into two steps: 1) finding the set of patterns that can be \((\varepsilon,\delta)\)-covered by others; 2) finding minimal P-RFPs by solving a set cover problem. The approach’s bottleneck is checking whether a pattern \((\varepsilon,\delta)\)-covers another in the first step, which involves the computation of a joint probability of the supports of two patterns. To address the problem, we propose a dynamic programming based approach, which iteratively updates the \((\varepsilon,\delta)\)-cover probability of two patterns in the first \( j \) transactions in an uncertain database. We also develop a set of effective optimization strategies to further improve the computation efficiency of the proposed approach.

To our knowledge, this is the first work that summarizes frequent patterns mined over uncertain databases by probabilistic representative pattern mining. Our experimental results show that our approach summarizes frequent patterns effectively, and restores the patterns and their original frequency probability information with a guaranteed error bound.

The remainder of the paper is structured as follows. The next section introduces related works to this paper. We define important concepts and provide a problem statement in Section 3. Section 4 describes the proposed data mining approach. Experimental results are presented in Section 5. Section 6 closes this paper with some conclusive remarks.

2 Related Work
In this section, we review related research from two sub-areas: frequent pattern mining over uncertain data and frequent pattern summarization.

Frequent pattern mining over uncertain data. Many approaches have been proposed to mine frequent patterns from uncertain databases in past years. Based on the definition of a frequent pattern, existing work on mining frequent patterns over uncertain data falls into two categories: expected support-based frequent pattern mining [3, 10, 11] and probabilistic frequent pattern mining [4, 5]. The former employs the expectation of support as the measurement. That is, a pattern is frequent only if its expected support is no less than a specified minimum expected support. The latter considers the frequency probability as the measurement, which refers to the probability that a pattern appears no less than a specified minimum number of support times. A pattern is therefore frequent only if its frequency probability is no less than a specified minimum probability (i.e. \( \text{Pr}(\text{supp}(X) \geq \text{minsup}) \geq \text{minprob} \)).

For mining expected support-based frequent patterns, there are three representative algorithms: UApriori [3], UFP-growth [10], UH-Mine [11]. UApriori is the uncertain version of the well-known Apriori algorithm. Both UFP-growth and UH-Mine are based on the divide-and-conquer framework that uses the depth-first strategy to search frequent patterns. For mining probabilistic frequent patterns, two representative algorithms are DP – dynamic programming-based Apriori algorithm [4], and DC – divide-and-conquer-based Apriori algorithm [5]. Observing that the support of a pattern in an uncertain database can be represented by poisson binomial distribution, some approximate probabilistic frequent pattern mining algorithms have also been proposed. [12, 13] respectively use the normal distribution and the poisson method to approximate the frequency probability of patterns. Recently, Tong et al. [6] verified that the two types of frequent patterns mined from uncertain data have a tight connection and can be unified when the size of data is large enough. They also empirically compared the performance of eight existing representative algorithms with uniform measures.

Frequent pattern summarization. Motivated by the fact that frequent pattern mining may generate an exponential number of patterns due to the downward closure property, a lot of research work has been dedicated to summarizing the complete set of patterns with a small set of representative ones. Various concepts have been proposed, such as maximal patterns [14], frequent closed patterns [15], and non-derivable patterns [16]. While all frequent patterns can be recov-
covered by

\(X\)

is the set of transactions supporting pattern

Definition 3.2.

to summarize frequent patterns mined over uncertain

databases based on their distance in possible worlds.

There are several generalizations of closed patterns, such as the pattern profiling based approaches [17, 18, 19] and the support distance based approaches [20, 21]. It was observed in [21] that the profile-based approaches [17, 18] have some drawbacks, such as no error guarantee on restored support. This work borrows the framework of the support distance based approaches to find probabilistic representative frequent patterns.

Recently, some research work has been undertaken to summarize frequent patterns mined over uncertain data. Tang and Peterson [8] proposed mining probabilistic frequent closed patterns, based on the concept called probabilistic support. Tong et al. [9] pointed out that frequent closed patterns defined on probabilistic support cannot guarantee the patterns are closed in possible worlds which contribute to their probabilistic supports. Instead, they defined the threshold-based frequent closed patterns over probabilistic data, which considers the probabilities of possible worlds where a pattern is closed. Our research relaxes the condition to further reduce the size of patterns by considering the probabilities of possible worlds where a pattern can \(\varepsilon\)-cover another one.

3 Problem Definition

This section first introduces preliminary definitions and then formulates the problem of probabilistic representative frequent pattern (P-RFP) mining.

Xin et al. [20] defined a robust distance measure between patterns in deterministic data.

Definition 3.1. (distance measure) Given two patterns \(X_1\) and \(X_2\), the distance between them, denoted as \(d(X_1, X_2)\), is defined as \(1 - \frac{|T(X_1) \cap T(X_2)|}{|T(X_1)|} \), where \(T(X_i)\) is the set of transactions supporting pattern \(X_i\).

Then, an \(\varepsilon\)-covered relationship is defined on two patterns where one subsumes another.

Definition 3.2. (\(\varepsilon\)-covered) Given a real number \(\varepsilon \in [0, 1]\) and two patterns \(X_1\) and \(X_2\), we say \(X_1\) is \(\varepsilon\)-covered by \(X_2\) if \(X_1 \subseteq X_2\) and \(d(X_1, X_2) \leq \varepsilon\).

As commonly used in frequent pattern mining, \(X_1 \subseteq X_2\) denotes \(X_1\) is a subset of \(X_2\) (e.g. \(\{a\} \subseteq \{a, b\}\)). It can be proved easily that, if \(X_2\) \(\varepsilon\)-covers \(X_1\), \(\frac{|\text{supp}(X_1) - \text{supp}(X_2)|}{\text{supp}(X_1)} \leq \varepsilon\). The goal of representative frequent pattern mining then becomes finding the minimal set of patterns that \(\varepsilon\)-cover all frequent patterns.

Table 1: An uncertain database with attribute uncertainty

<table>
<thead>
<tr>
<th>ID</th>
<th>Transactions</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(w_1)</td>
<td>({T_1 : \phi, T_2 : {a}})</td>
<td>0.12</td>
</tr>
<tr>
<td>(w_2)</td>
<td>({T_1 : {a}, T_2 : {a}})</td>
<td>0.28</td>
</tr>
<tr>
<td>(w_3)</td>
<td>({T_1 : {b}, T_2 : {a}})</td>
<td>0.03</td>
</tr>
<tr>
<td>(w_4)</td>
<td>({T_1 : {a, b}, T_2 : {a}})</td>
<td>0.07</td>
</tr>
<tr>
<td>(w_5)</td>
<td>({T_1 : \phi, T_2 : {a, c}})</td>
<td>0.12</td>
</tr>
<tr>
<td>(w_6)</td>
<td>({T_1 : {a}, T_2 : {a, c}})</td>
<td>0.28</td>
</tr>
<tr>
<td>(w_7)</td>
<td>({T_1 : {b}, T_2 : {a, c}})</td>
<td>0.03</td>
</tr>
<tr>
<td>(w_8)</td>
<td>({T_1 : {a, b}, T_2 : {a, c}})</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 2: An example of possible worlds

In the context of uncertain data, the support of a pattern, \(\text{supp}(X_i)\), becomes a discrete random variable. Therefore, we cannot directly apply the \(\varepsilon\)-covered relationship to probabilistic frequent patterns. Before explaining how to extend the concept of \(\varepsilon\)-covered in the context of uncertain data, we examine an uncertain database where attributes are associated with existential probabilities. Table 1 shows an uncertain transaction database where each transaction consists of a set of probabilistic items. For example, the probability that item \(a\) appears in the first transaction \(T_1\) is 0.7. Possible world semantics are commonly used to explain the existence of data in an uncertain database. For example, the database in Table 1 has eight possible worlds, which are listed in Table 2. Each possible world is associated with an existential probability. For instance, the probability that the first possible world \(w_1\) exists is \((1 - 0.7) \times (1 - 0.2) \times 1 \times (1 - 0.5) = 0.12\).

Considering that the occurrences of items in every possible world are deterministic, we can define the probabilistic distance between two probabilistic frequent patterns based on their distance in possible worlds.

Definition 3.3. (probabilistic distance measure) Given an uncertain database \(D\), and two patterns \(X_1\) and \(X_2\), let \(\mathcal{PW} = \{w_1, \ldots, w_n\}\) be the set of possible worlds derived from \(D\), the distance between \(X_1\) and \(X_2\) in a possible world \(w_j\) in \(\mathcal{PW}\) is

\[
(3.1) \quad \text{dist}(X_1, X_2; w_j) = 1 - \frac{|T(X_1; w_j) \cap T(X_2; w_j)|}{|T(X_1; w_j) \cup T(X_2; w_j)|}
\]

where \(T(X_i; w_j)\) is the set of transactions containing pattern \(X_i\) in the possible world \(w_j\). Then, the probabilistic distance between \(X_1\) and \(X_2\), denoted by
dist(X₁, X₂), is a random variable. The probability mass function of dist(X₁, X₂) is:
\[
\Pr(\text{dist}(X₁, X₂) = d) = \sum_{wₖ \in PW, \text{dist}(X₁, X₂; wₖ) = d} \Pr(wₖ)
\]

That is, the probability that the distance between two probabilistic frequent patterns is d equals to the sum of the probabilities of possible worlds where the distance between the two patterns is d.

For example, consider the uncertain database in Table 1. Let X₁ = \{a\} and X₂ = \{a, b\}. The probability that the distance between X₁ and X₂ is equal to 0.5, \( \Pr(\text{dist}(X₁, X₂) = 0.5) \), can be computed by adding the probabilities of the possible worlds \( w₄ \) and \( w₈ \). This is because only in the two possible worlds, the distance between the two patterns is 0.5. Therefore, \( \Pr(\text{dist}(X₁, X₂) = 0.5) = 0.14 \).

Based on the probabilistic distance measure, we define the \( \varepsilon \)-cover probability as follows.

**Definition 3.4.** (\( \varepsilon \)-cover probability) Given an uncertain database \( D \), two patterns \( X₁ \) and \( X₂ \), and a distance threshold \( \varepsilon \), the \( \varepsilon \)-cover probability of \( X₁ \) and \( X₂ \) is \( \Pr_{\text{cover}}(X₁, X₂; \varepsilon) = \Pr(\text{dist}(X₁, X₂) \leq \varepsilon) \).

**Definition 3.5.** (\( \varepsilon, \delta \)-covered) Given an uncertain database \( D \), two patterns \( X₁ \) and \( X₂ \), a distance threshold \( \varepsilon \) and a cover probability threshold \( \delta \), we say \( X₂ \) \( \varepsilon, \delta \)-covers \( X₁ \) if \( X₁ \subseteq X₂ \) and \( \Pr_{\text{cover}}(X₁, X₂; \varepsilon) \geq \delta \).

Our goal is then to obtain the minimal set of patterns that will \( \varepsilon, \delta \)-cover all the probabilistic frequent patterns. The formal statement of the probabilistic representative frequent pattern (P-RFP) mining is as follows.

**Definition 3.6.** (Problem Statement) Given an uncertain database \( D \), a set of probabilistic frequent patterns \( F \), a probabilistic distance threshold \( \varepsilon \) and a cover probability threshold \( \delta \), the problem of probabilistic representative frequent pattern (P-RFP) mining is to find the minimal set of patterns \( \mathcal{R} \) so that, for any frequent pattern \( X \in \mathcal{F} \), there exists a representative pattern \( X' \in \mathcal{R} \) where \( X' \varepsilon, \delta \)-covers \( X \).

It is obvious that when \( \varepsilon = 0 \), the probabilistic representative pattern set is probabilistic closed patterns, and when \( \varepsilon = 1 \), it is probabilistic maximal patterns.

### 4 P-RFP Mining

This section first describes the framework of our proposed approach. Then, we explain the details of the main steps for P-RFP mining.

#### 4.1 Framework of P-RFP Mining

Before presenting the framework of our approach for P-RFP mining, we develop some important lemmas between two patterns where one \( \varepsilon, \delta \)-covers another.

**Lemma 4.1.** Given an uncertain database \( D \) and two patterns \( X₁ \) and \( X₂ \) s.t. \( X₂ \varepsilon, \delta \)-covers \( X₁ \), the distance between \( X₁ \) and \( X₂ \) in the possible world \( wₗ \) can be represented by the support of the patterns in \( wₗ \):
\[
\text{dist}(X₁, X₂; wₗ) = 1 - \frac{\text{supp}(X₂; wₗ)}{\text{supp}(X₁; wₗ)}
\]

**Proof.** Since \( X₂ \varepsilon, \delta \)-covers \( X₁ \), then \( X₁ \subseteq X₂ \),
\[
\text{dist}(X₁, X₂; wₗ) = 1 - \frac{|T(X₁; wₗ)\cap T(X₂; wₗ)|}{|T(X₁; wₗ)|} = 1 - \frac{|T(X₂; wₗ)|}{|T(X₁; wₗ)|} = 1 - \frac{\text{supp}(X₂; wₗ)}{\text{supp}(X₁; wₗ)} \tag{4.3}
\]

**Lemma 4.2.** Given an uncertain database \( D \) and two patterns \( X₁ \) and \( X₂ \) s.t. \( X₂ \varepsilon, \delta \)-covers \( X₁ \), the probabilistic distance \( \text{dist}(X₁, X₂) \) can be represented by the support distribution of \( X₁ \) and \( X₂ \):
\[
\text{dist}(X₁, X₂) = 1 - \frac{\text{supp}(X₂)}{\text{supp}(X₁)}
\]

**Proof.** \( \text{supp}(Xᵢ) \) is a discrete random variable. \( \Pr(\text{supp}(Xᵢ) = k) = \sum_{wᵢ \in PW, \text{supp}(Xᵢ; wᵢ) = k} \Pr(wᵢ) \). According to the definition of probabilistic distance and Lemma 4.1, we have
\[
\Pr(\text{dist}(X₁, X₂) = d) = \sum_{wᵢ \in PW, \text{dist}(X₁, X₂; wᵢ) = d} \Pr(wᵢ)
\]
\[
= \sum_{wᵢ \in PW, 1 - \frac{\text{supp}(X₂; wᵢ)}{\text{supp}(X₁; wᵢ)} = d} \Pr(wᵢ)
\]
For brevity, \( W' = \{wᵢ | wᵢ \in PW, \text{supp}(X₂; wᵢ) = k, \text{supp}(X₁; wᵢ) = (1 - d)k\} \), where \( k \in [0, |D|] \), then
\[
\Pr(\text{dist}(X₁, X₂) = d) = \sum_{|D|} \sum_{k=1} \Pr(wᵢ)
\]
\[
= \sum_{k=1} \Pr(\text{supp}(X₂) = k, \text{supp}(X₁) = (1 - d)k)
\]
\[
= \Pr((1 - \frac{\text{supp}(X₂)}{\text{supp}(X₁)}) = d) \tag{4.5}
\]

**Lemma 4.3.** Given an uncertain database \( D \) and two patterns \( X₁ \) and \( X₂ \) s.t. \( X₂ \varepsilon, \delta \)-covers \( X₁ \), we have
\[
\Pr(\text{supp}(X₂) \geq (1 - \varepsilon) \text{supp}(X₁)) \geq \delta
\]
Proof. Since $X_2 (\varepsilon, \delta)$-covers $X_1$, according to Lemma 4.2, we have

$$
\Pr(\text{dist}(X_1, X_2) \leq \varepsilon) = \Pr\left(1 - \frac{\text{supp}(X_2)}{\text{supp}(X_1)} \leq \varepsilon\right)
$$

$$
= \Pr(\text{supp}(X_2) \geq (1 - \varepsilon)\text{supp}(X_1)) \geq \delta
$$

Hence, $Pr(\text{dist}(X_1, X_2) \leq \varepsilon) \geq \delta$. According to Lemma 4.3, the $\varepsilon$-cover probability $Pr_{\text{cover}}(X_1, X_2; \varepsilon) = Pr(\text{dist}(X_1, X_2) \leq \varepsilon)$ is equivalent to $Pr(\text{supp}(X_2) \geq (1 - \varepsilon)\text{supp}(X_1))$. Let $\text{supp}(X_1) = l$, and $(\text{supp}(X_2) = k)$. Then, the $\varepsilon$-cover probability between $X_1$ and $X_2$ is equal to,

$$
\sum_{l=0}^{\lfloor |D| \rfloor} \sum_{k=\lfloor (1-\varepsilon)l \rfloor}^{l} \Pr(\text{supp}(X_1) = l, \text{supp}(X_2) = k)
$$

To compute the value of Equation (4.6) to find out whether it is no less than $\delta$, we introduce the joint support probability distribution as follows.

**Definition 4.1.** (joint support probability) Given an uncertain database $D$ and patterns $X_1$ and $X_2$, the joint support probability mass function is

$$
\Pr(\text{supp}(X_1) = l, \text{supp}(X_2) = k) = \sum_{w_i \in \mathcal{PW}, \text{supp}(X_1; w_i) = l, \text{supp}(X_2; w_i) = k} \Pr(w_i)
$$

To split the computation of the joint support probability of $X_1$ and $X_2$ into smaller sub-problems, we define the partial joint support probability distribution as follows.

**Definition 4.2.** (partial joint support probability) Given an uncertain database $D$ and patterns $X_1$ and $X_2$, the $j$-partial joint support probability is the joint support probability of $X_1$ and $X_2$ in the first $j$ transactions of $D$. The $j$-partial joint support mass function is

$$
\sum_{w_i \in \mathcal{PW}, \text{supp}(X_1; w_i) = l, \text{supp}(X_2; w_i) = k} \Pr(w_i)
$$

It can be proved that the partial joint support probability can be computed in a recursive strategy.

**Lemma 4.5.** Given an uncertain database $D$ and patterns $X_1$ and $X_2$, $X_1 \subseteq X_2$, $j, k, l \in \mathbb{Z}$ and $0 \leq k \leq l \leq
Table 3: Probability of different situations in the jth transaction

<table>
<thead>
<tr>
<th>Situation</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1 \subseteq t_j$, $X_2 \subseteq t_j$</td>
<td>$p_j^{X_1}$</td>
</tr>
<tr>
<td>$X_1 \subseteq t_j$, $X_2 \not\subseteq t_j$</td>
<td>$p_j^{X_1} - p_j^{X_2}$</td>
</tr>
<tr>
<td>$X_1 \not\subseteq t_j$, $X_2 \not\subseteq t_j$</td>
<td>$1 - p_j^{X_1}$</td>
</tr>
</tbody>
</table>

Lemma 4.7. Given an uncertain database $D$ and two patterns $X_1$ and $X_2$ s.t. $X_1 \subseteq X_2$, if $X_2 (\varepsilon, \delta)$-covers $X_1$, then $\forall X$ s.t. $X_1 \subseteq X \subseteq X_2$, we have $X_2 (\varepsilon, \delta)$-covers $X$.

Proof. Since $X_2 (\varepsilon, \delta)$-covers $X_1$, according to Lemma 4.3 we have

$$\Pr(\text{supp}(X_2) \geq (1 - \varepsilon)\text{supp}(X_1)) \geq \delta$$

$\forall X$ that $X_1 \subseteq X \subseteq X_2$, we have $\text{supp}(X_1; w_j) \geq \text{supp}(X; w_j) \geq \text{supp}(X_2; w_j)$ in every possible world $w_j$. Therefore, $\Pr(\text{supp}(X_2) \geq (1 - \varepsilon)\text{supp}(X_1)) \geq \delta \Rightarrow \Pr(\text{supp}(X_2) \geq (1 - \varepsilon)\text{supp}(X_1)) \geq (1 - \varepsilon)\text{supp}(X) \geq \delta$

Hence, $X_2 (\varepsilon, \delta)$-covers $X$. □

According to Lemma 4.7, we have the following corollary.

Corollary 4.1. Given an uncertain database $D$ and two patterns $X_1$ and $X_2$, $X_1 \subseteq X_2$, if $X_2$ cannot $(\varepsilon, \delta)$-cover $X_1$, then $\forall X \subseteq X_1$, $X_2$ cannot $(\varepsilon, \delta)$-cover $X$.

Lemma 4.7 and Corollary 4.1 reduce the number of pattern pairs, for which the cover probability needs to be computed.

Lemma 4.8. Given an uncertain database $D$, two patterns $X_1$ and $X_2$ s.t. $X_1 \subseteq X_2$, a distance threshold $\varepsilon$, and a cover probability threshold $\delta$, if $\exists j, 1 \leq j \leq |D|$ such that $\prod_{m=1}^{j} (1 - p_m^{X_1} + p_m^{X_2}) \geq \delta$, then $X_2 (\varepsilon, \delta)$-covers $X_1$, where $p_m^{X_i}$ is the probability that $X_i$ occurs in the m-th transaction.

Proof. Recall that, according to Equation (4.6), the $\varepsilon$-cover probability $\Pr(\text{dist}(X_1, X_2) \leq \varepsilon)$ is equivalent to,

$$\sum_{l=0}^{|D|} \sum_{k=\lceil(1-\varepsilon)l\rceil}^{l} \Pr(\text{supp}(X_1) = l, \text{supp}(X_2) = k)$$

which is greater than $Q(|D|) = \sum_{l=0}^{|D|} \Pr(\text{supp}(X_1) = \text{supp}(X_2) = l)$. Therefore, if $Q(|D|) \geq \delta$, then $\Pr(\text{dist}(X_1, X_2) \leq \varepsilon) \geq \delta$. The computation of $Q(|D|)$ can be similarly split into smaller problems by introducing the partial $Q$ in the first $j$ transactions of $D$, $Q(j) = \sum_{l=0}^{j} \Pr(\text{supp}(X_1) = \text{supp}(X_2) = l)$. $Q(j)$ can
then be iteratively updated as follows (the details of the inference are provided in the supplementary document),

\[ Q(j) = Q(j - 1)(1 - p_j^{X_1} + p_j^{X_2}) \]
\[ = Q(j - 2)(1 - p_{j-1}^{X_1} + p_{j-1}^{X_2})(1 - p_j^{X_1} + p_j^{X_2}) \]
\[ \cdots \]
\[ = \prod_{m=1}^{j} (1 - p_m^{X_1} + p_m^{X_2}) \]

Hence, if \( \prod_{m=1}^{j} (1 - p_m^{X_1} + p_m^{X_2}) \geq \delta \), then \( X_2 (\varepsilon, \delta) \)-covers \( X_1 \). \( \square \)

Lemma 4.8 defines a lower bound of \( \varepsilon \)-cover probability, which can be computed efficiently using continued multiplication. If the lower bound is less than the cover probability threshold \( \delta \), we can immediately decide \( X_2 \) cannot \( (\varepsilon, \delta) \)-cover \( X_1 \) without computing their real \( \varepsilon \)-cover probability.

4.4 P-RFP Mining Algorithm. The overall framework of our P-RFP mining algorithm is shown in Algorithm 1. From lines 3 – 9, we find the cover set for each pattern \( X_2 \) in the pseudo probabilistic frequent patterns \( F \). The most important step is to check whether \( X_2 \) covers \( X_1 \in F \) (line 6). The details of the function \( isCover \) is illustrated in Algorithm 2, where lines 1 – 3 implement the optimization stated by Lemma 4.7, and lines 4 – 6 apply the Corollary 4.1. Lines 7 – 9 in the function \( isCover \) use Lemma 4.8 to efficiently discover the lower bound of \( \varepsilon \)-cover probability. Note that, according to Lemma 4.6, the lower bound, as well as the real \( \varepsilon \)-cover probability, can be computed on a sub-database \( D(X_1) \). Finally, from lines 10 – 14, we use the dynamic programming based scheme to compute the \( \varepsilon \)-cover probability. As mentioned before, the function \( setCover \) in Algorithm 1 is solved using the greedy algorithm in [22].

Algorithm 1 P-RFP-Mining Framework

Input: \( D, F, \hat{F}, \varepsilon \) and \( \delta \)
Output: Minimal P-RFP Set \( R \)
1: \( R \leftarrow \Phi \)
2: \( CoverSets \leftarrow \Phi \)
3: for all \( X_2 \in \hat{F} \) do
4: \( NoCoverSet \leftarrow \Phi \)
5: for all \( X_1 \in F \) such that \( X_1 \subseteq X_2 \) do
6: \( \text{if } isCover(X_1, X_2) = True \text{ then } \)
7: \( CoverSets[X_2].add(X_1) \)
8: \( \text{else } \)
9: \( NoCoverSet.add(X_1) \)
10: \( R = setCover(CoverSets, F) \)
11: return \( R \)

Algorithm 2 Function \( isCover \)

Input: \( X_1, \hat{X}_2 \)
Output: If \( X_2 (\varepsilon, \delta) \)-covers \( X_1 \), then return True, else False
1: for all \( X \in CoverSets[X_2] \) do
2: \( \text{if } X \subseteq X_1 \text{ then } \)
3: \( \text{return True } \)
4: for all \( X \in NoCoverSet[X_2] \) do
5: \( \text{if } X \supseteq X_1 \text{ then } \)
6: \( \text{return False } \)
7: \( Q(|D(X_1)|) = \prod_{m=1}^{D(X_1)} (1 - p_m^{X_1} + p_m^{X_2}) \)
8: \( \text{if } Q(|D(X_1)|) \geq \delta \text{ then } \)
9: \( \text{return True } \)
10: \( \text{for } l = 0 \text{ to } |D(X_1)| \text{ do } \)
11: \( \text{for } k = [(1 - \varepsilon)l] \text{ to } l - 1 \text{ do } \)
12: \( P_{cover} = \Pr(supp(X_1) = l, supp(X_2) = k) \)
13: \( \text{if } P_{cover} \geq \delta \text{ then } \)
14: \( \text{return True } \)
15: return False

5 Performance Study
This section evaluates the effectiveness of P-RFPs, the performance of our approach for P-RFP mining, and the optimization strategies.

5.1 Data sets. Two datasets have been used in our experiments. The first is the Retail dataset from the Frequent Itemset Mining (FIMI) Dataset Repository. This is one of the standard datasets used in frequent pattern mining in deterministic databases. In order to bring uncertainty into the dataset, we synthesize an existential probability for each item based on a Gaussian distribution with the mean of 0.9 and the variance of 0.125. This dataset is an uncertain database that associates uncertainty to attributes.

The second one is the iceberg sighting record from 1993 to 1997 on the North Atlantic from the International Ice Patrol (IIP) Iceberg Sightings Database. The IIP Iceberg Sighting Database collects information of iceberg activities in the North Atlantic. Each transaction in the database contains the information of date, location, size, shape, reporting source and a confidence level. The confidence level has six possible attributes, R/V (Radar and visual), R (Radar only), V (Visual), MEA (Measured), EST (Estimated) and GBL (Garbled), which indicate different reliabilities of that tuple. We translate confidence levels to probabilities 0.8, 0.7, 0.6, 0.5, 0.4 and 0.3, respectively. This dataset is an uncertain database that associates uncertainty to tuples.
5.2 Result analysis. We first evaluate the compression rate of the P-RFPs, with respect to the variation of parameters. We randomly select 1000 transactions from the two datasets respectively to conduct the experiment. The sizes of \( R \) - the set of P-RFPs, and \( F \) - the set of probabilistic frequent patterns, with respect to the variations of \( \text{minsup} \), \( \text{minprob} \), \( \epsilon \), and \( \delta \), on the two datasets are shown in Figures 1 and 2 respectively. The default values of the four parameters are set to 0.5%, 0.8, 0.2 and 0.2 respectively. It can be observed from the results on both datasets, when \( \text{minsup} \) and \( \text{minprob} \) are low, the compression rate of P-RFPs is high because there are more probabilistic frequent patterns. For the variations of \( \epsilon \) and \( \delta \), obviously, the high compression rate can be achieved if the probabilistic distance threshold \( \epsilon \) is high and/or the cover probability threshold \( \delta \) is low.

We then examine the runtime of the proposed algorithm for P-RFP mining. Figures 3 and 4 show the runtime vs. \( \text{minsup} \), \( \text{minprob} \), \( \epsilon \), and \( \delta \) curves on 1000 transactions randomly selected from the two datasets, respectively. The default values of the four parameters are same as in the first experiment. It is intuitive that, when \( \epsilon \) is increasing or \( \text{minsup} \), \( \text{minprob} \) and \( \delta \) are decreasing, the runtime will increase because more pattern pairs are engaged in cover probability checking. We find that the growth of both \( \epsilon \) and \( \delta \) lead to a tradeoff between the number of P-RFPs and runtime.

We also evaluate the effectiveness of the optimization strategies proposed in sub-section 4.3. We randomly select 500 transactions from the two datasets, respectively, to carry out this experiment. The default values for the experiments on the Retail dataset are: \( \text{minsup} = 4\% \), \( \text{minprob} = 0.8 \), \( \epsilon = 0.1 \) and \( \delta = 0.2 \). On the IIP dataset, the four parameters are set to 10\%, 0.8, 0.1 and 0.2 by default, respectively. Figure 5 shows the runtime of the basic version of our algorithm, and the runtime of the algorithm integrated with optimization strategies, with respect to the variation of \( \epsilon \) and \( \delta \) on the two datasets, respectively. The results clearly reveal the effectiveness of the optimization strategies by demonstrating that the optimized algorithm significantly reduces the runtime.
6 Conclusions
Due to the downward closure property, the number of probabilistic frequent patterns mined over uncertain data can be so large that they hinder further analysis and exploitation. This paper proposes the P-RFP mining, which aims to find a small set of patterns to represent the complete set of probabilistic frequent patterns. To address the data uncertainty issue, we define the concept of probabilistic distance, as well as a $(\varepsilon, \delta)$-cover relationship between two patterns. P-RFPs are the minimal set of patterns that $(\varepsilon, \delta)$-cover the complete set of probabilistic frequent patterns. We develop a P-RFP mining algorithm that uses a dynamic programming based scheme to efficiently check whether one pattern $(\varepsilon, \delta)$-covers another. We also exploit effective optimization strategies to further improve the computation efficiency. Our experimental results demonstrate that the devised data mining algorithm effectively and efficiently discovers the set of P-RFPs, which can substantially reduce the size of probabilistic frequent patterns.

This work extends the measure defined in deterministic databases to quantify the distance between two patterns in terms of their supporting transactions. Since the supports of patterns are random variables in the context of uncertain data, other distance measures, such as Kullback-Leibler divergence, might be applicable. As ongoing work, we will study the effectiveness of probabilistic representative frequent patterns defined on different distance measures.

7 Acknowledgements
This work was supported, in part, by the Australian Research Council (ARC) Linkage Project under Grant No. LP120100566.

References