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

[*] QCIS, University of Technology, Sydney Email: Chunyang.Liu@student.uts.edu.au

[†] QCIS, University of Technology, Sydney Email: Ling.Chen@uts.edu.au

[‡] QCIS, University of Technology, Sydney Email: Chengqi.Zhang@uts.edu.au

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 ε-*covered* relationship between patterns. A pattern X_1 is ε-covered by another pattern X_2 if X_1 is a subset of X_2 and $(supp(X_1) - supp(X_2))/supp(X_1) \leq \varepsilon$. The goal is then to find a minimum set of representative patterns that can ε-cover all frequent patterns. Since the support of a pattern $(supp(X))$ becomes a discrete random variable in an uncertain database, the ε-covered relationship cannot be applied directly to probabilistic frequent patterns.

In this work, we first extend the concept of ε-covered by defining a new (ε, δ)-*covered* relationship between probabilistic frequent patterns. Informally, a pattern X_1 is (ε, δ)-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 ε is no less than δ. The objective of probabilistic representative frequent pattern $(P-RFP)$ mining is then to find the minimal set of patterns that can (ε, δ)-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 (ε, δ)-covered by others; 2) finding minimal P-RFPs by solving a set cover problem. The approach's bottleneck is checking whether a pattern (ε, δ)-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 (ε, δ)-*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 subareas: 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. $Pr(supp(X) \geq minsup) \geq minprob$).

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 recovered from maximal patterns, the support information is lost. Although the set of frequent closed patterns preserve the exact support of all frequent patterns, the number of frequent closed patterns can still be tens of thousands or even more. 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 ε-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 $1 - \frac{|T(X_1) \cap T(X_2)|}{|T(X_1) \cup T(X_2)|}$, where $T(X_i)$ is the set of transactions supporting pattern X_i.*

Then, an ε-covered relationship is defined on two patterns where one subsumes another.

Definition 3.2. (ε*-covered*) *Given a real number $\varepsilon \in [0, 1]$ and two patterns X_1 and X_2, we say X_1 is ε-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 ε-covers X_1, $\frac{supp(X_1) - supp(X_2)}{supp(X_1)} \leq \varepsilon$. The goal of representative frequent pattern mining then becomes finding the minimal set of patterns that ε-cover all frequent patterns.

ID	Transactions
T_1	a:0.7 b:0.2
T_2	a:1.0 c:0.5

Table 1: An uncertain database with attribute uncertainty

ID	Possible World	Prob.
w_1	$\{T_1 : \phi, T_2 : \{a\}\}$	0.12
w_2	$\{T_1 : \{a\}, T_2 : \{a\}\}$	0.28
w_3	$\{T_1 : \{b\}, T_2 : \{a\}\}$	0.03
w_4	$\{T_1 : \{a, b\}, T_2 : \{a\}\}$	0.07
w_5	$\{T_1 : \phi, T_2 : \{a, c\}\}$	0.12
w_6	$\{T_1 : \{a\}, T_2 : \{a, c\}\}$	0.28
w_7	$\{T_1 : \{b\}, T_2 : \{a, c\}\}$	0.03
w_8	$\{T_1 : \{a, b\}, T_2 : \{a, c\}\}$	0.07

Table 2: An example of possible worlds

In the context of uncertain data, the support of a pattern, $supp(X_i)$, becomes a discrete random variable. Therefore, we cannot directly apply the ε-cover relationship to probabilistic frequent patterns. Before explaining how to extend the concept of ε-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 $\mathcal{PW}= \{w_1, \ldots, w_m\}$ 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*

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

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

$dist(X_1, X_2)$, is a random variable. The probability mass function of $dist(X_1, X_2)$ is:

(3.2)
$$\Pr(dist(X_1, X_2) = d) = \sum_{w_j \in \mathcal{PW}, dist(X_1, X_2; w_j) = d} \Pr(w_j)$$

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 $X_1 = \{a\}$ and $X_2 = \{a, b\}$. The probability that the distance between X_1 and X_2 is equal to 0.5, $Pr(dist(X_1, X_2) = 0.5)$, can be computed by adding the probabilities of the possible worlds w_4 and w_8. This is because only in the two possible worlds, the distance between the two patterns is 0.5. Therefore, $Pr(dist(X_1, X_2) = 0.5) = 0.14$.

Based on the probabilistic distance measure, we define the ε-cover probability as follows.

Definition 3.4. (ε*-cover probability*) *Given an uncertain database D, two patterns X_1 and X_2, and a distance threshold ε, the ε-cover probability of X_1 and X_2 is $\Pr_{cover}(X_1, X_2; \varepsilon) = \Pr(dist(X_1, X_2) \leq \varepsilon)$.*

Definition 3.5. ((ε, δ)*-covered*) *Given an uncertain database D, two patterns X_1 and X_2, a distance threshold ε and a cover probability threshold δ, we say X_2 (ε, δ)-covers X_1 if $X_1 \subseteq X_2$ and $\Pr_{cover}(X_1, X_2; \varepsilon) \geq \delta$.*

Our goal is then to obtain the minimal set of patterns that will (ε, δ)-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 $\mathcal{F}$, a probabilistic distance threshold ε and a cover probability threshold δ, 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' (ε, δ)-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 (ε, δ)-covers another.

Lemma 4.1. *Given an uncertain database D and two patterns X_1 and X_2 s.t. X_2 (ε, δ)-covers X_1, the distance between X_1 and X_2 in the possible world w_j can be represented by the support of the patterns in w_j:*

(4.3)
$$dist(X_1, X_2; w_j) = 1 - \frac{supp(X_2; w_j)}{supp(X_1; w_j)}$$

Proof. Since X_2 (ε, δ)-covers X_1, then $X_1 \subseteq X_2$,

$$dist(X_1, X_2; w_j) = 1 - \frac{|T(X_1; w_j) \cap T(X_2; w_j)|}{|T(X_1; w_j) \cup T(X_2; w_j)|}$$
$$= 1 - \frac{|T(X_2; w_j)|}{|T(X_1; w_j)|} = 1 - \frac{supp(X_2; w_j)}{supp(X_1; w_j)} \qquad \square$$

Lemma 4.2. *Given an uncertain database D and two patterns X_1 and X_2 s.t. X_2 (ε, δ)-covers X_1, the probabilistic distance $dist(X_1, X_2)$ can be represented by the support distribution of X_1 and X_2:*

(4.4)
$$dist(X_1, X_2) = 1 - \frac{supp(X_2)}{supp(X_1)}$$

Proof. $supp(X_i)$ is a discrete random variable. $\Pr(supp(X_i) = k) = \sum_{w_j \in \mathcal{PW}, supp(X; w_j) = k} \Pr(w_j)$. According to the definition of probabilistic distance and Lemma 4.1, we have

$$\Pr(dist(X_1, X_2) = d) = \sum_{w_j \in \mathcal{PW}, dist(X_1, X_2; w_j) = d} \Pr(w_j)$$
$$= \sum_{w_j \in \mathcal{PW}, 1 - \frac{supp(X_2; w_j)}{supp(X_1; w_j)} = d} \Pr(w_j)$$

For brevity, le $W' = \{w_j \, | w_j \in \mathcal{PW}, supp(X_2; w_j) = k, supp(X_1; w_j) = (1-d)k\}$, where $k \in [0, |D|]$, then

$$\Pr(dist(X_1, X_2) = d) = \sum_{k=1}^{|D|} \sum_{w_j \in W'} \Pr(w_j)$$
$$= \sum_{k=1}^{|D|} \Pr(supp(X_2) = k, supp(X_1) = (1-d)k)$$
$$= \Pr((1 - \frac{supp(X_2)}{supp(X_1)}) = d) \qquad \square$$

Lemma 4.3. *Given an uncertain database D and two patterns X_1 and X_2 s.t. X_2 (ε, δ)-covers X_1, we have*

(4.5)
$$\Pr\left(supp(X_2) \geq (1 - \varepsilon) supp(X_1)\right) \geq \delta$$

Proof. Since X_2 (ε,δ)-covers X_1, according to Lemma 4.2, we have

$$\Pr(dist(X_1,X_2)\le\varepsilon)=\Pr\left(1-\frac{supp(X_2)}{supp(X_1)}\le\varepsilon\right)$$
$$=\Pr(supp(X_2)\ge(1-\varepsilon)supp(X_1))\ge\delta \qquad \square$$

LEMMA 4.4. *Given an uncertain database D, two patterns X_1 and X_2, a support threshold minsup and a frequency probability threshold minprob, if X_2 (ε,δ)-covers X_1, and X_1 is a probabilistic frequent pattern w.r.t. minsup and minprob, then X_2 is a probabilistic frequent pattern w.r.t. $(1-\varepsilon)minsup$ and $(\delta\cdot minprob)$.*

Proof. Since X_1 is a probabilistic frequent pattern w.r.t. *minsup* and *minprob*, we have $\Pr(supp(X_1)\ge minsup)\ge minprob$, which infers,

$$\Pr((1-\varepsilon)supp(X_1)\ge(1-\varepsilon)minsup))\ge minprob$$

From Lemma 4.3, we have,

$$\Pr(supp(X_2)\ge(1-\varepsilon)supp(X_1))\ge\delta$$

Hence, $\Pr(supp(X_2)\ge(1-\varepsilon)minsup)\ge\delta\cdot minprob$. That is, X_2 is a probabilistic frequent pattern w.r.t. $((1-\varepsilon)minsup)$ and $(\delta\cdot minprob)$. $\square$

According to Lemma 4.4, to find the representative patterns to (ε,δ)-cover the complete set of probabilistic frequent patterns w.r.t. *minsup* and *minprob*, denoted as F, we need to consider the set of pseudo probabilistic frequent patterns w.r.t $(1-\varepsilon)minsup$ and $(\delta\cdot minprob)$, denoted as $\hat{F}$. Given the two sets F and $\hat{F}$, our approach for P-RFP mining consists of the following two steps.

1. Generate the cover set for every pattern in $\hat{F}$. For each pattern X in $\hat{F}$, the cover set of X, denoted as $C(X)$, is a set of probabilistic frequent patterns in F that can be (ε,δ)-covered by X. That is, $C(X)\subseteq F$.

2. Find the minimal pattern set $R\subseteq\hat{F}$ to (ε,δ)-cover all probabilistic frequent patterns in F.

After finding the cover sets for patterns in $\hat{F}$ in the first step, the second step is equivalent to finding a minimal number of cover sets that cover all patterns in F. This is known as a set cover problem, which is NP-hard. Similar to [21], we adopt a well-known greedy set cover algorithm [22], which achieves polynomial complexity. Therefore, in the following, we focus on describing the first step, which generates the cover set for each pseudo probabilistic frequent pattern in $\hat{F}$.

4.2 Cover Set Generation.

To generate the cover set for a pattern X_2 in $\hat{F}$, for each pattern X_1 in F so that $X_1\subseteq X_2$, we need to check if X_2 (ε,δ)-covers X_1. That is, we need to examine whether the ε-cover probability between X_1 and X_2 is no less than δ (i.e., $\Pr(dist(X_1,X_2)\le\varepsilon)\ge\delta$). According to Lemma 4.3, the ε-cover probability $\Pr_{cover}(X_1,X_2;\varepsilon)$=$\Pr(dist(X_1,X_2)\le\varepsilon)$ is equivalent to $\Pr(supp(X_2)\ge(1-\varepsilon)supp(X_1))$. Let $supp(X_1)=l$, and $supp(X_2)=k$. Then, the ε-cover probability between X_1 and X_2 is equal to,

$$(4.6)\quad \sum_{l=0}^{|D|}\sum_{k=\lceil(1-\varepsilon)l\rceil}^{l}\Pr(supp(X_1)=l,supp(X_2)=k))$$

To compute the value of Equation (4.6) to find out whether it is no less than δ, 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*

$$(4.7)\qquad \Pr(supp(X_1)=l,supp(X_2)=k)$$
$$=\sum_{w_i\in\mathcal{PW},supp(X_1;w_i)=l,supp(X_2;w_i)=k}\Pr(w_i)$$

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*

$$(4.8)\qquad \Pr_j(supp(X_1)=l,supp(X_2)=k)$$
$$=\sum_{w_i\in\mathcal{PW},supp_j(X_1;w_i)=l,supp_j(X_2;w_i)=k}\Pr(w_i)$$

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 $0\le k\le l\le$*

Situation	Probability
$X_1 \subseteq t_j$, $X_2 \subseteq t_j$	$p_j^{X_2}$
$X_1 \subseteq t_j$, $X_2 \nsubseteq t_j$	$p_j^{X_1} - p_j^{X_2}$
$X_1 \nsubseteq t_j$, $X_2 \nsubseteq t_j$	$1 - p_j^{X_1}$

Table 3: Probability of different situations in the jth transaction

$j \leq |D|$, then:

(4.9)

$$\begin{aligned}
&\Pr_j(supp(X_1) = l, supp(X_2) = k) \\
&= \Pr_{j-1}(supp(X_1) = l, supp(X_2) = k)(1 - p_j^{X_1}) \\
&+ \Pr_{j-1}(supp(X_1) = l - 1, supp(X_2) = k)(p_j^{X_1} - p_j^{X_2}) \\
&+ \Pr_{j-1}(supp(X_1) = l - 1, supp(X_2) = k - 1)p_j^{X_2}
\end{aligned}$$

where $p_j^{X_i}$ is the probability that X_i occurs in the jth transaction. The boundary case is: $\Pr_j(supp(X_1) = 0, supp(X_2) = 0) = \prod_{m=1}^{j}(1 - p_m^{X_1})$.

Proof. In jth-transaction $t_j \in D$, there are only three existence possibilities of X_1 and X_2, since $X_1 \subseteq X_2$. Table 3 lists the three situations and their respective existential probabilities. Therefore, we can split the computation of $\Pr_j(supp(X_1) = l, supp(X_2) = k)$ into the three situations with corresponding probability. The equation of the boundary case is intuitive. □

Lemma 4.5 enables us to compute the cover probability iteratively using a dynamic programming scheme. This equation is the foundation of our approach. Although it is feasible and effective, we can accelerate it through certain optimization techniques, which are stated in the next sub-section.

4.3 Optimization Strategies. While Lemma 4.5 approves the cover probability between two patterns can be updated transaction by transaction, the transactions which support neither of the two patterns can actually be skipped.

LEMMA 4.6. *Given an uncertain database D, two patterns X_1 and X_2 s.t. $X_1 \subseteq X_2$, and a probabilistic distance threshold ε, $\Pr_{cover}(X_1, X_2; \varepsilon)$ computed on D is equal to that computed on $D(X_1)$, where $D(X_1)$ is $\{t | P(X_1 \subseteq t) > 0, t \in D\} \subseteq D$.*

Lemma 4.6 is intuitive. According to Definition 3.3 and Definition 3.4, only the transactions supporting at least the sub-pattern X_1 will contribute to the value of probabilistic distance, which in turn affects the ε-cover probability. This lemma allows us to compute the ε-cover probability on a projected sub-database, which significantly reduces the runtime of computation.

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 (ε, δ)-covers X_1, then $\forall X$ s.t. $X_1 \subseteq X \subseteq X_2$, we have X_2 (ε, δ)-covers X.*

Proof. Since X_2 (ε, δ)-covers X_1, according to Lemma 4.3 we have

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

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

$$\Pr(supp(X_2) \geq (1 - \varepsilon)supp(X_1) \geq (1 - \varepsilon)supp(X)) \geq \delta$$

Hence, X_2 (ε, δ)-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 (ε, δ)-cover X_1, then $\forall X \subseteq X_1$, X_2 cannot (ε, δ)-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 ε, and a cover probability threshold δ, 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 (ε, δ)-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 ε-cover probability $\Pr(dist(X_1, X_2) \leq \varepsilon)$ is equivalent to,

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

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

$$\begin{aligned} Q(j) &= Q(j-1)(1-p_j^{X_1}+p_j^{X_2}) \\ &= Q(j-2)(1-p_{j-1}^{X_1}+p_{j-1}^{X_2})(1-p_j^{X_1}+p_j^{X_2}) \\ &\dots \\ &= \prod_{m=1}^{j}(1-p_m^{X_1}+p_m^{X_2}) \end{aligned}$$

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

Lemma 4.8 defines a lower bound of ε-cover probability, which can be computed efficiently using continued multiplication. If the lower bound is less than the cover probability threshold δ, we can immediately decide X_2 cannot (ε, δ)-cover X_1 without computing their real ε-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 $\hat{F}$. 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 ε-cover probability. Note that, according to Lemma 4.6, the lower bound, as well as the real ε-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 ε-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:         if $isCover(X_1, X_2) = True$ then
 7:             $CoverSets[X_2].add(X_1)$
 8:         else
 9:             $NoCoverSet.add(X_1)$
10: $R = setCover(CoverSets, F)$
11: return $R$
```

Algorithm 2 Function *isCover*

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

[1] http://fimi.cs.helsinki.fi/data/

[2] http://nsidc.org/data/g00807.html

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 $minsup$, $minprob$, ε, and δ, 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 $minsup$ and $minprob$ are low, the compression rate of P-RFPs is high because there are more probabilistic frequent patterns. For the variations of ε and δ, obviously, the high compression rate can be achieved if the probabilistic distance threshold ε is high and/or the cover probability threshold δ is low.

We then examine the runtime of the proposed algorithm for P-RFP mining. Figures 3 and 4 show the runtime vs. $minsup$, $minprob$, ε, and δ 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 ε is increasing or $minsup$, $minprob$ and δ are decreasing, the runtime will increase because more pattern pairs are engaged in cover probability checking. We find that the growth of both ε and δ lead to a tradeoff between the number of P-RFPs and runtime.

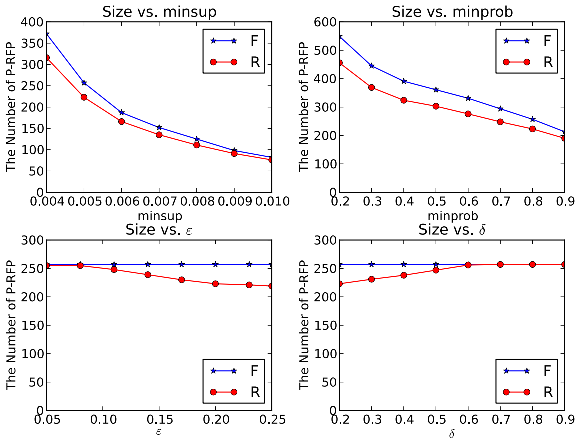

Figure 1: The Number of P-RFP on Retail

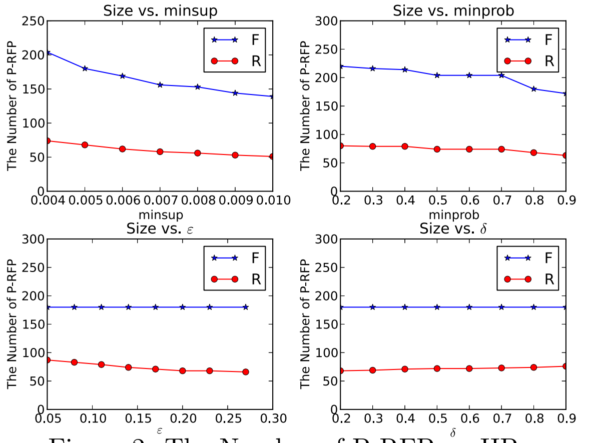

Figure 2: The Number of P-RFP on IIP

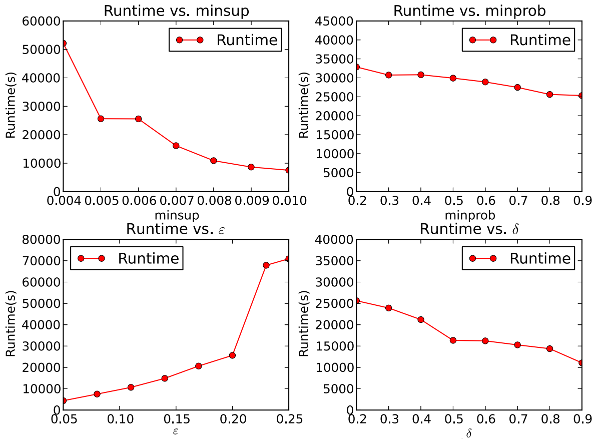

Figure 3: Runtime on Retail

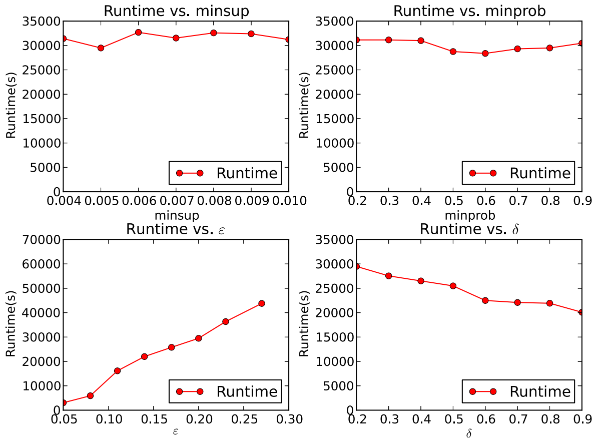

Figure 4: Runtime on IIP

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: $minsup$ = 4%, $minprob$ = 0.8, ε = 0.1 and δ = 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 ε and δ 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.

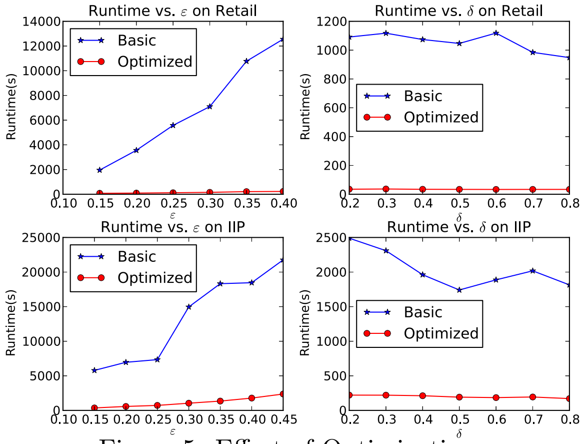

Figure 5: Effect of Optimization

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 (ε, δ)-cover relationship between two patterns. P-RFPs are the minimal set of patterns that (ε, δ)-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 (ε, δ)-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

[1] Aggarwal, C.C., Yu, P.S.: A survey of uncertain data algorithms and applications. IEEE Transactions on Knowledge and Data Engineering **21**(5) (2009) 609–623

[2] Aggarwal, C.C.: Managing and mining uncertain data. Springer (2009)

[3] Chui, C.K., Kao, B., Hung, E.: Mining frequent itemsets from uncertain data. PAKDD (2007) 47–58

[4] Bernecker, T., Kriegel, H.P., Renz, M., Verhein, F., Zuefle, A.: Probabilistic frequent itemset mining in uncertain databases. SIGKDD (2009) 119–128

[5] Sun, L., Cheng, R., Cheung, D.W., Cheng, J.: Mining uncertain data with probabilistic guarantees. SIGKDD (2010) 273–282

[6] Tong, Y., Chen, L., Cheng, Y., Yu, P.S.: Mining frequent itemsets over uncertain databases. VLDB Endowment **5**(11) (2012) 1650–1661

[7] Peterson, E.A., Tang, P.: Fast approximation of probabilistic frequent closed itemsets. ASRC (2012) 214–219

[8] Tang, P., Peterson, E.A.: Mining probabilistic frequent closed itemsets in uncertain databases. ASRC (2011) 86–91

[9] Tong, Y., Chen, L., Ding, B.: Discovering threshold-based frequent closed itemsets over probabilistic data. ICDE (2012) 270–281

[10] Leung, C., Mateo, M., Brajczuk, D.: A tree-based approach for frequent pattern mining from uncertain data. Advances in Knowledge Discovery and Data Mining (2008) 653–661

[11] Aggarwal, C.C., Li, Y., Wang, J.: Frequent pattern mining with uncertain data. SIGKDD (2009) 29–38

[12] Calders, T., Garboni, C., Goethals, B.: Approximation of frequentness probability of itemsets in uncertain data. ICDE (2010) 749–754

[13] Wang, L., Cheng, R., Lee, S.D., Cheung, D.: Accelerating probabilistic frequent itemset mining: a model-based approach. CIKM (2010) 429–438

[14] Bayardo Jr., R. J.: Efficiently mining long patterns from databases. SIGMOD (1998) 85–93

[15] Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. ICDT (1999) 398–416

[16] Calders, T., Goethals, B.: Mining all non-derivable frequent itemsets. PKDD (2002) 74–85

[17] Yan, X., Cheng, H., Han, J., Xin, D.: Summarizing itemset patterns: a profile-based approach. SIGKDD (2005) 314–323

[18] Jin, R., Abu-Ata, M., Xiang, Y., Ruan, N.: Effective and efficient itemset pattern summarization: regression-based approaches. SIGKDD (2008) 399–407

[19] Poernomo, A.K., Gopalkrishnan, V.: Cp-summary: a concise representation for browsing frequent itemsets. SIGKDD (2009) 687–696

[20] Xin, D., Han, J., Yan, X., Cheng, H.: Mining compressed frequent-pattern sets. VLDB (2005) 709–720

[21] Liu, G., Zhang, H., Wong, L.: Finding minimum representative pattern sets. SIGKDD (2012) 51–59

[22] Chvatal, V.: A greedy heuristic for the set-covering problem. Mathematics of operations research **4**(3) (1979) 233–235

Renew SIAM · Contact Us · Site Map · Join SIAM · My Account
Society for Industrial and Applied Mathema s
Search this site:
search

SIAM Home
About SIAM
Activity Groups
Advertising
Books
Careers & Jobs
Conferences
Customer Service
Digital Library
Fellows Program
History Project
Journals
Membership
Prizes & Recognitions
Proceedings
Public Awareness
Reports
Sections
SIAM Connect
SIAM News
Students
Visiting Lecturer
Find us on Facebook
FOLLOW US ON twitter

CONFERENCES > SIAM INTERNATIONAL CONFERENCE ON DATA MINING >

Paper Submission and Key Dates

Deadline Dates

September 30, 2012 11:59 PM PST:	Workshop/Tutorial Proposals
~~October 12, 2012~~ October 15, 2012 11:59 PM PST:	Paper Submission
December 20, 2012:	Author Notification
January 30, 2013:	Camera Ready Papers Due

Paper Submission Instructions

Papers submitted to this conference must not have been accepted or be under review by another conference with a published proceedings or by a journal. The work may be either theoretical or applied, but should make a significant contribution to the field.

All papers should have a maximum length of 9 pages with US Letter (8.5" x 11") paper size (single-spaced, 2 column, 10 point font, and at least 1" margin on each side). Papers must have an abstract with a maximum of 300 words and a keyword list with no more than 6 keywords. Authors are required to submit their papers electronically in PDF format to https://cmt.research.microsoft.com/SDM2013/ by 11:59 PM (PST),~~October 12, 2012~~ **October 15, 2012**. Postscript files can be converted using standard converters.

In addition, authors will be given an opportunity to upload a supplementary file (see instructions below) with maximum length of 5 pages to CMT after submitting the paper itself. Both the paper and supplementary file must be in PDF format, and the maximum file size for each file is 10MB.

We would like to encourage you to prepare your paper in LaTeX2e. Papers should be formatted using the SIAM SODA macro, which is available through the SIAM website. You can access it at http://www.siam.org/proceedings/macros.php . The filename is **soda2e.all**. Make sure you use the macros for SODA and Data Mining Proceedings; papers prepared using other proceedings macros will not be accepted.

For Microsoft Word users, please convert your document to the PDF format. Since there is no Microsoft Word Template, please visit http://www.siam.org/proceedings/ to view the format on previous papers.

All submissions should clearly present the author information including the names of the authors, the affiliations and the emails.

In This Section

Home
Abstract Download
Child Care
Exhibits
General Information
Hotel Information
Invited Presentations
Navigating a SIAM Conference
Other Data Mining Meetings
Paper Submission and Key Dates
Proceedings
Program and Abstracts
Program Updates
Register Now
Registration Information
Related Links
SIAM Presents
Sponsors
Students
Titles on Display
Travel Support
Tutorials
Workshops

Submission Site

Visit the SDM13 Conference Submission Site at https://cmt.research.microsoft.com/SDM2013/default.aspx to submit. At this website you will be able to upload your PDF file and select a Primary Subject Area for your paper. You should also select as many Secondary Areas as your paper fits.

These selections are important to the review process and must be done with care. Note that papers may be reallocated to a more appropriate subject area if need be at the discretion of the program chairs.

Paper Structure

In SDM13, in addition to traditional work related to the design, analysis, and implementation of data mining algorithms and systems we also strongly encourage submissions of an applied nature. To facilitate the fair and effective reviewing of such application-oriented submissions, we encourage authors to structure such papers into the following sections.

1. Background and Motivation (Specific to the application domain)
2. Methods and Technical Solutions(Draw connections between the specific application to existing studies in the literature, clarify the constraints imposed by the application domain and distinguish your solutions from pre-existing ones in this context).
3. Empirical Evaluation (Self explanatory).
4. Significance and Impact (Provide concrete evidence of the potential significance and impact of your work in the given application domain).

The best papers in SDM13 will be invited to be expanded and submitted to a special issue in the Journal of Statistical Analysis and Data Mining.

Instructions for Submitting Optional Supplementary Materials

All submissions must be in PDF format, and must follow the SIAM style files. Papers are limited to 9 pages; however, authors should feel free to send submissions that are shorter than 9 pages. This includes the title, author details, abstract, technical details, empirical results and bibliography.

In addition, authors have the *option* of submitting a supplementary file for their paper (to present details such as long proofs, additional experimental results, implementation or system details, etc.). The supplementary file must be in PDF format, follow the SIAM style files, and must not exceed 5 pages. To submit the supplementary file, please submit the main paper first, and then you will be given an opportunity to upload the supplementary file to the CMT system.

Please note the following:

- Submissions with the main body longer than 9 pages, or supplementary file longer than 5 pages, will be rejected without review.
- Only the main body (9 pages) of the submissions of accepted papers will be published in SDM proceedings. However, authors of accepted papers are encouraged to publicize their supplementary files on their own websites.
- It is up to the discretion of the PCs and SPCs to look at the supplementary files, and the SPCs and PCs reserve the right to judge the quality of the paper solely on the basis of the 9 pages.

Conflict of Interest Guidelines for Submissions

As part of the submission procedure authors are asked to mark conflicts of interest with Program Committee members.

A paper author has a conflict of interest with a Program Committee member if any of the following hold:

1. The Program Committee member is an advisee or advisor of any one of the authors. This applies to current and former advisees and advisors.
2. The Program Committee member is a collaborator or co-author within the last two years of one of the authors. Collaborations include things like joint papers published or in submission as well as joint projects either in progress or within the last two years.
3. The Program Committee member is a relative or close personal friend of one of the authors.
4. The Program Committee member is part of the same organization or has been a part of the same organization as one of the authors within the last two years.

Inaccurate representation of conflicts of interest can result in the paper being rejected without review at the discretion of the program chairs.

If you believe there are other conditions causing a Program Committee member to have a conflict of interest with your submitted paper please contact the Program Committee chairs at: sdm2013chairs@gmail.com.

Authorial Integrity in Scientific Publication

SIAM Policies and Procedures on Authorial Integrity

3600 Market Street, 6th Floor | Philadelphia, PA 19104-2688 USA
Phone: +1-215-382-9800 | FAX: +1-215-386-7999
site map | privacy policy | webmaster@siam.org | suggestions
Banner art adapted from a figure by Hinke M. Osinga and Bernd Krauskopf (University of Auckland, NZ).