k-Anonymous Patterns

Maurizio Atzori[1,2], Francesco Bonchi[2], Fosca Giannotti[2], and Dino Pedreschi[1]

[1] Pisa KDD Laboratory, Computer Science Department, University of Pisa, Italy
e-mail: {atzori, pedre}@di.unipi.it
[2] Pisa KDD Laboratory, ISTI - CNR, Pisa, Italy
e-mail: {francesco.bonchi, fosca.giannotti}@isti.cnr.it

Abstract. It is generally believed that data mining results do not violate the *anonymity* of the individuals recorded in the source database. In fact, data mining models and patterns, in order to ensure a required statistical significance, represent a large number of individuals and thus conceal individual identities: this is the case of the minimum support threshold in association rule mining. In this paper we show that this belief is ill-founded. By shifting the concept of *k-anonymity* from data to patterns, we formally characterize the notion of a threat to anonymity in the context of pattern discovery, and provide a methodology to efficiently and effectively identify all possible such threats that might arise from the disclosure of a set of extracted patterns.

1 Introduction

Privacy Preserving Data Mining, i.e., the analysis of data mining side-effects on privacy, has recently become a key research issue and is receiving a growing attention from the research community [1, 3, 9, 16]. However, despite such efforts, a common understanding of what is meant by "privacy" is still missing. This fact has led to the proliferation of many completely different approaches to privacy preserving data mining, all sharing the same generic goal: producing a valid mining model without disclosing "private" data. As highlighted in [9], the approaches pursued so far leave a privacy question open: do the data mining results themselves violate privacy? Put in other words, do the disclosure of extracted patterns open up the risk of privacy breaches that may reveal sensitive information? During the last year, few works [7, 9, 11] have tried to address this problem by some different points of view, but they all require some *a priori* knowledge of what is sensitive and what is not.

In this paper we study when data mining results represent *per se* a threat to privacy, independently of any background knowledge of what is sensitive. In particular, we focus on *individual privacy*, which is concerned with the *anonymity* of individuals.

A prototypical application instance is in the medical domain, where the collected data are typically very sensitive, and the kind of privacy usually required is the anonymity of the patients in a survey. Consider a medical institution where the usual hospital activity is coupled with medical research activity. Since physicians are the data collectors and holders, and they already know everything about their patients, they have unrestricted access to the collected information. Therefore, they can perform real mining on all available information using traditional mining tools – not necessarily the privacy

preserving ones. This way they maximize the outcome of the knowledge discovery process, without any concern about privacy of the patients which are recorded in the data. But the anonymity of patients becomes a key issue when the physicians want to share their discoveries (e.g., association rules holding in the data) with their scientific community.

At a first sight, it seems that data mining results do not violate the anonymity of the individuals recorded in the source database. In fact, data mining models and patterns, in order to ensure a required statistical significance, represent a large number of individuals and thus conceal individual identities: this is the case of the minimum support threshold in association rule mining. The next example shows that the above belief is ill-founded.

Example 1. Consider the following association rule:

$$a_1 \wedge a_2 \wedge a_3 \Rightarrow a_4 \quad [sup = 80,\ conf = 98.7\%]$$

where sup and $conf$ are the usual interestingness measures of *support* and *confidence* as defined in [2]. Since the given rule holds for a number of individuals (80), which seems large enough to protect individual privacy, one could conclude that the given rule can be safely disclosed. But, is this all the information contained in such a rule? Indeed, one can easily derive the support of the premise of the rule:

$$sup(\{a_1, a_2, a_3\}) = \frac{sup(\{a_1, a_2, a_3, a_4\})}{conf} \approx \frac{80}{0.987} = 81.05$$

Given that the pattern $a_1 \wedge a_2 \wedge a_3 \wedge a_4$ holds for 80 individuals, and that the pattern $a_1 \wedge a_2 \wedge a_3$ holds for 81 individuals, we can infer that in our database there is just one individual for which the pattern $a_1 \wedge a_2 \wedge a_3 \wedge \neg a_4$ holds.

The knowledge inferred is a clear threat to the anonymity of that individual: on one hand the pattern identifying the individual could itself contain sensitive information; on the other hand it could be used to re-identify the same individual in other databases.

It is worth noting that this problem is very general: the given rule could be, instead of an association, a classification rule, or the path from the root to the leaf in a decision tree, and the same reasoning would still hold. Moreover, it is straightforward to note that, unluckily, the more accurate is a rule, the more unsafe it may be w.r.t. anonymity. As shown later, this anonymity problem can not be simply solved by discarding the most accurate rules: in fact, more complex kinds of threats to anonymity exist which involve more than simply two itemsets.

1.1 Related Works

During the last years a novel problem has emerged in privacy-preserving data mining [7, 9, 11]: do the data mining results themselves violate privacy? Only little preliminary work is available. The work in [9] studies the case of a classifier trained over a mixture of different kind of data: *public* (known to every one including the adversary), *private/sensitive* (should remain unknown to the adversary), and *unknown* (neither sensitive nor known by the adversary). The authors propose a model for privacy implication of the learned classifier.

In [11] the data owner, rather than sharing the data, prefers to share the mined association rules, but requires that a set of *restricted* association rules are not disclosed. The authors propose a framework to sanitize the output of association rules mining, while blocking some inference channels for the restricted rules.

In [7] a framework for evaluating classification rules in terms of their perceived privacy and ethical sensitivity is described. The proposed framework empowers the data miner with alerts for sensitive rules that can be accepted or dismissed by the user as appropriate. Such alerts are based on an aggregate *sensitivity combination function*, which assigns to each rule a value of sensitivity by aggregating the sensitivity value (an integer between 0 and 9) of each attribute involved in the rule. The process of labelling each attribute with its sensitivity value must be accomplished by the domain expert.

The fundamental difference of these approaches with ours lies in generality: we propose a novel, objective definition of privacy compliance of patterns without any reference to a preconceived knowledge of sensitive data or patterns, on the basis of the rather intuitive and realistic constraint that the anonymity of individuals should be guaranteed.

An important method for protecting individual privacy is k-*anonymity*, introduced in [14], a notion that establishes that the cardinality of the answer to any possible query should be at least k. In this work, it is shown that protection of individual sources does not guarantee protection when sources are cross-examined: a sensitive medical record, for instance, can be uniquely linked to a *named* voter record in a publicly available voter list through some shared attributes. The objective of k-anonymity is to eliminate such opportunities of inferring private information through cross linkage. In particular, this is obtained by a "sanitization" of the source data that is transformed in such a way that, for all possible queries, at least k tuples will be returned. Such a sanitization is obtained by generalization and suppression of attributes and/or tuples [15].

Trivially, by mining a k-anonymized database no patterns threatening the anonymity can be obtained. But such mining would produce models impoverished by the information loss which is intrinsic in the generalization and suppression techniques. Since our objective is to extract valid and interesting patterns, we propose to postpone k-anonymization after the actual mining step. In other words, we do not to enforce k-anonymity onto the source data, but instead we move such a concept to the extracted patterns.

1.2 Paper Contributions

In this paper we study the privacy problem described above in the very general setting of patterns which are boolean formulas over a binary database. Our contribution is twofold:

- we define k-anonymous patterns and provide a general characterization of inference channels holding among patterns that may threat anonymity of source data;
- we develop an effective and efficient algorithm to detect such potential threats, which yields a methodology to check whether the mining results may be disclosed without any risk of violating anonymity.

We emphasize that the capability of detecting the potential threats is extremely useful for the analyst to determine a trade-off among the quality of mining result and the privacy guarantee, by means of an iterative interaction with the proposed detection algorithm. Our empirical experiments, reported in this paper, bring evidence to this observation. It should also be noted the different setting w.r.t. the other works in privacy preserving data mining: in our context no data perturbation or sanitization is performed; we allow real mining on the real data, while focussing on the *anonymity preservation properties of the extracted patterns*. We have also developed possible strategies to eliminate the threats to anonymity by introducing distortion on the dangerous patterns in a controlled way: for lack of space these results are omitted here but can be found in [5].

2 k-Anonymous Patterns and σ-Frequent Itemsets

We start by defining binary databases and patterns following the notation in [8].

Definition 2. A binary database $\mathcal{D} = (\mathcal{I}, \mathcal{T})$ consists of a finite set of binary variables $\mathcal{I} = \{i_1, \ldots, i_p\}$, also known as *items*, and a finite multiset $\mathcal{T} = \{t_1, \ldots, t_n\}$ of p-dimensional binary vectors recording the values of the items. Such vectors are also known as *transactions*. A *pattern* for the variables in $\mathcal{I}$ is a logical (propositional) sentence built by AND ($\wedge$), OR ($\vee$) and NOT ($\neg$) logical connectives, on variables in $\mathcal{I}$. The domain of all possible patterns is denoted $\mathcal{P}at(\mathcal{I})$.

According to Def. 2, $e \wedge (\neg b \vee \neg d)$, where $b, d, e \in \mathcal{I}$, is a pattern. One of the most important properties of a pattern is its frequency in the database, i.e. the number of individuals (transactions) in the database which make the given pattern true[1].

Definition 3. Given a database $\mathcal{D}$, a transaction $t \in \mathcal{D}$ and a pattern p, we write $p(t)$ if t makes p true. The *support* of p in $\mathcal{D}$ is given by the number of transactions which makes p true: $sup_{\mathcal{D}}(p) = |\{t \in \mathcal{D} \mid p(t)\}|$.

The most studied *pattern class* is the itemset, i.e., a conjunction of positive valued variables, or in other words, a set of items. The retrieval of itemsets which satisfy a minimum frequency property is the basic step of many data mining tasks, including (but not limited to) association rules [2, 4].

Definition 4. The set of all *itemsets* $2^{\mathcal{I}}$, is a pattern class consisting of all possible conjunctions of the form $i_1 \wedge i_2 \wedge \ldots \wedge i_m$. Given a database $\mathcal{D}$ and a minimum support threshold σ, the set of *σ-frequent itemsets* in $\mathcal{D}$ is denoted

$$\mathcal{F}(\mathcal{D}, \sigma) = \{\langle X, sup_{\mathcal{D}}(X)\rangle \mid X \in 2^{\mathcal{I}} \wedge sup_{\mathcal{D}}(X) \geq \sigma\}$$

Itemsets are usually denoted in the form of set of the items in the conjunction, e.g. $\{i_1, \ldots, i_m\}$; or sometimes, simply $i_1 \ldots i_m$. Figure 1(b) shows the different notation used for general patterns and for itemsets. The problem addressed in this paper is given

[1] The notion of truth of a pattern w.r.t. a transaction t is defined in the usual way: t makes p true iff t is a model of the propositional sentence p.

$\mathcal{D}$

	a	b	c	d	e	f	g	h
t_1	1	1	1	1	1	1	1	1
t_2	1	1	1	1	1	0	1	0
t_3	1	1	1	1	1	0	0	0
t_4	1	1	1	1	1	1	1	0
t_5	1	1	1	1	1	0	0	0
t_6	1	1	1	1	1	0	0	0
t_7	1	1	0	1	1	0	0	0
t_8	1	0	0	0	1	1	1	0
t_9	0	0	1	1	1	1	1	0
t_{10}	0	0	1	1	1	0	0	0
t_{11}	0	0	1	1	1	1	1	1
t_{12}	1	1	0	0	0	1	1	0

(a)

Notation: patterns

$sup_{\mathcal{D}}(a \vee f) = 11$
$sup_{\mathcal{D}}(e \wedge (\neg b \vee \neg d)) = 4$
$sup_{\mathcal{D}}(h \wedge \neg b) = 1$

Notation: itemsets

$sup_{\mathcal{D}}(abc) = 6$
$sup_{\mathcal{D}}(abde) = 7$
$sup_{\mathcal{D}}(cd) = 9$

(b)

$\mathcal{F}(\mathcal{D}, 8) = \{\langle \emptyset, 12\rangle, \langle a, 9\rangle, \langle b, 8\rangle, \langle c, 9\rangle, \langle d, 10\rangle, \langle e, 11\rangle, \langle ab, 8\rangle, \langle ae, 8\rangle, \langle cd, 9\rangle, \langle ce, 9\rangle, \langle de, 10\rangle, \langle cde, 9\rangle\}$

(c)

$\mathcal{C}l(\mathcal{D}, 8) = \{\langle \emptyset, 12\rangle, \langle a, 9\rangle, \langle e, 11\rangle, \langle ab, 8\rangle, \langle ae, 8\rangle, \langle de, 10\rangle, \langle cde, 9\rangle\}$

(d)

$\mathcal{MC}h(3, \mathcal{C}l(\mathcal{D}, 6)) = \{\langle \mathcal{C}_{abde}^{abcde}, 1\rangle, \langle \mathcal{C}_{ae}^{abcde}, 1\rangle, \langle \mathcal{C}_{ab}^{abcde}, 1\rangle, \langle \mathcal{C}_{g}^{fg}, 1\rangle, \langle \mathcal{C}_{g}^{eg}, 1\rangle\}$

(e)

Fig. 1. Running example: (a) the binary database $\mathcal{D}$; (b) different notation used for patterns and itemsets; (c) the set of σ-frequent ($\sigma = 8$) itemsets; (d) the set of closed frequent itemsets; (e) the set of maximal inference channels for $k = 3$ and $\sigma = 6$.

by the possibility of inferring from the output of frequent itemset mining, i.e, $\mathcal{F}(\mathcal{D}, \sigma)$, the existence of patterns with very low support (i.e., smaller than an anonymity threshold k, but not null): such patterns represent a threat for the anonymity of the individuals about which they are true.

Definition 5. Given a database $\mathcal{D}$ and an anonymity threshold k, a pattern p is said to be *k-anonymous* if $sup_{\mathcal{D}}(p) \geq k$ or $sup_{\mathcal{D}}(p) = 0$.

2.1 Problem Definition

Before introducing our anonymity preservation problem, we need to define the inference of supports, which is the basic tool for the attacks to anonymity.

Definition 6. A set S of pairs $\langle X, n\rangle$, where $X \in 2^{\mathcal{I}}$ and $n \in \mathbb{N}$, and a database $\mathcal{D}$ are said to be σ-compatible if $S = \mathcal{F}(\mathcal{D}, \sigma)$. Given a pattern p we say that $S \models sup(p) > x$ (respectively $S \models sup(p) < x$) if, for all databases $\mathcal{D}$ σ-compatible with S, we have that $sup_{\mathcal{D}}(p) > x$ (respectively $sup_{\mathcal{D}}(p) < x$).

Informally, we call *inference channel* any subset of the collection of itemsets (with their respective supports), from which it is possible to infer non k-anonymous patterns. Our mining problem can be seen as frequent pattern extraction with two frequency thresholds: the usual minimum support threshold σ for itemsets (as defined in Definition 4), and an anonymity threshold k for general patterns (as defined in Definition 2).

Note that an itemset with support less than k is itself a non k-anonymous, and thus dangerous, pattern. However, since we can safely assume (as we will do in the rest of this paper) that $\sigma \gg k$, such pattern would be discarded by the usual mining algorithms.

Definition 7. Given a collection of frequent itemsets $\mathcal{F}(\mathcal{D}, \sigma)$ and an anonymity threshold k, our problem consists in detecting all possible inference channels $\mathcal{C} \subseteq \mathcal{F}(\mathcal{D}, \sigma)$: $\exists p \in \mathcal{P}at(\mathcal{I}) : \mathcal{C} \models 0 < sup_{\mathcal{D}}(p) < k$.

Obviously, a solution to this problem directly yields a method to formally prove that the disclosure of a given collection of frequent itemsets does not violate the anonymity constraint: it is sufficient to check that no inference channel exists for the given collection. In this case, the collection can be safely distributed even to malicious adversaries. On the contrary, if this is not the case, we can proceed in two ways:

- mine a new collection of frequent itemsets under different circumstances, e.g., higher minimum support threshold, to look for an admissible collection;
- transform (sanitize) the collection to remove the inference channels.

The second alternative opens up many interesting mining problems, which are omitted here for lack of space, and are discussed in [5].

3 Detecting Inference Channels

In this Section we study how information about non k-anonymous patterns can be possibly inferred from a collection of σ-frequent itemsets. As suggested by Example 1, a simple inference channel is given by any itemset X which has a superset $X \cup \{a\}$ such that $0 < sup_{\mathcal{D}}(X) - sup_{\mathcal{D}}(X \cup \{a\}) < k$. In this case the pair $\langle X, sup_{\mathcal{D}}(X)\rangle, \langle X \cup \{a\}, sup_{\mathcal{D}}(X \cup \{a\})\rangle$ is an inference channel for the non k-anonymous pattern $X \wedge \neg a$, whose support is directly given by $sup_{\mathcal{D}}(X) - sup_{\mathcal{D}}(X \cup \{a\})$. This is a trivial kind of inference channel. Do more complex structures of itemsets exist that can be used as inference channels? In general, the support of a pattern $p = i_1 \wedge \cdots \wedge i_m \wedge \neg a_1 \wedge \cdots \wedge \neg a_n$ can be inferred if we know the support of itemsets $I = \{i_1, \ldots, i_m\}$, $J = I \cup \{a_1, \ldots, a_n\}$, and every itemset L such that $I \subset L \subset J$.

Lemma 8. *Given a pattern $p = i_1 \wedge \cdots \wedge i_m \wedge \neg a_1 \wedge \cdots \wedge \neg a_n$ we have that:*

$$sup_{\mathcal{D}}(p) = \sum_{I \subseteq X \subseteq J} (-1)^{|X \setminus I|} sup_{\mathcal{D}}(X)$$

where $I = \{i_1, \ldots, i_m\}$ and $J = I \cup \{a_1, \ldots, a_n\}$.

Proof. (Sketch) The proof follows directly from the definition of support and the well-known *inclusion-exclusion principle* [10].

Following the notation in [6], we denote the right-hand side of the equation above as $f_I^J(\mathcal{D})$. In the database $\mathcal{D}$ in Figure 1 we have that $sup_{\mathcal{D}}(b \wedge \neg d \wedge \neg e) = f_b^{bde}(\mathcal{D}) = sup_{\mathcal{D}}(b) - sup_{\mathcal{D}}(bd) - sup_{\mathcal{D}}(be) + sup_{\mathcal{D}}(bde) = 8 - 7 - 7 + 7 = 1$.

Definition 9. Given a database $\mathcal{D}$, and two itemsets $I, J \in 2^{\mathcal{I}}$, $I = \{i_1, \ldots, i_m\}$ and $J = I \cup \{a_1, \ldots, a_n\}$, if $0 < f_I^J(\mathcal{D}) < k$, then the set of itemsets $\{X | I \subseteq X \subseteq J\}$ constitutes an inference channel for the non k-anonymous pattern $p = i_1 \wedge \cdots \wedge i_m \wedge \neg a_1 \wedge \cdots \wedge \neg a_n$. We denote such inference channel $\mathcal{C}_I^J$ and we write $sup_{\mathcal{D}}(\mathcal{C}_I^J) = f_I^J(\mathcal{D})$.

Example 10. Consider the database $\mathcal{D}$ of Figure 1, and suppose $k = 3$. We have that $\mathcal{C}_{ab}^{abcde}$ is an inference channel of support 1. This means that there is only one transaction $t \in \mathcal{D}$ is such that $a \wedge b \wedge \neg c \wedge \neg d \wedge \neg e$.

The next Theorem states that if there exists a non k-anonymous pattern, then there exists a pair of itemsets $I \subseteq J \in 2^{\mathcal{I}}$ such that $\mathcal{C}_I^J$ is an inference channel.

Theorem 11. $\forall p \in \mathcal{P}at(\mathcal{I}) : 0 < sup_{\mathcal{D}}(p) < k \;.\; \exists\, I \subseteq J \in 2^{\mathcal{I}} : \mathcal{C}_I^J$.

Proof. The case of a conjunctive pattern p is a direct consequence of Lemma 8. Let us now consider a generic pattern $p \in \mathcal{P}at(\mathcal{I})$. Without loss of generality p is in *normal disjunctive form*: $p = p_1 \vee \ldots \vee p_q$, where $p_1 \ldots p_q$ are conjunctive patterns. We have that:

$$sup_{\mathcal{D}}(p) \geq \max_{1 \leq i \leq q} sup_{\mathcal{D}}(p_i).$$

Since $sup_{\mathcal{D}}(p) < k$ we have for all patterns p_i that $sup_{\mathcal{D}}(p_i) < k$. Moreover, since $sup_{\mathcal{D}}(p) > 0$ is there at least a pattern p_i such that $sup_{\mathcal{D}}(p_i) > 0$. Therefore, there is at least a conjunctive pattern p_i such that $0 < sup_{\mathcal{D}}(p_i) < k$.

From Theorem 11 we conclude that all possible threats to anonymity are due to inference channels of the form $\mathcal{C}_I^J$. However we can divide such inference channels in two subgroups:

1. inference channels involving only frequent itemsets;
2. inference channels involving also infrequent itemsets.

The first problem, addressed in the rest of this paper, is the most essential. In fact, a malicious adversary can easily find inference channels made up only of elements which are present in the disclosed output. However, these inference channels are not the unique possible source of inference: further inference channels involving also infrequent itemsets could be possibly discovered, albeit in a much more complex way.

In fact, in [6] deduction rules to derive tight bounds on the support of itemsets are introduced. Given an itemset J, if for each subset $I \subset J$ the support $sup_{\mathcal{D}}(I)$ is known, such rules allow to compute lower and upper bounds on the support of J. Let l be the greatest lower bound we can derive, and u the smallest upper bound we can derive: if we find that $l = u$ then we can infer that $sup_{\mathcal{D}}(J) = l = u$ without actual counting. In this case J is said to be a *derivable itemset*. We transpose such deduction techniques in our context and observe that they can be exploited to discover information about infrequent itemsets, and from these to infer non k-anonymous patterns. For lack of space, this higher-order problem is not discussed here, and left to the extended version of this paper. However, here we can say that the techniques to detect this kind of inference channels and to block them are very similar to the techniques for the first kind of channels. This is due to the fact that both kinds of channels rely on the same concept: inferring supports of larger itemsets from smaller ones. Indeed, the key equation of our work (Lemma 8) is also the basis of the deduction rules proposed in [6].

From now on we restrict our attention to the essential form of inference channel, namely those involving frequent itemsets only.

Definition 12. The set of all $\mathcal{C}_I^J$ holding in $\mathcal{F}(\mathcal{D}, \sigma)$, together with their supports, is denoted $\mathcal{C}h(k, \mathcal{F}(\mathcal{D}, \sigma)) = \{\langle \mathcal{C}_I^J, f_I^J(\mathcal{D})\rangle \mid 0 < f_I^J(\mathcal{D}) < k \wedge \langle J, sup_{\mathcal{D}}(J)\rangle \in \mathcal{F}(\mathcal{D}, \sigma)\}$.

Algorithm 1 Naïve Inference Channel Detector

Input: $\mathcal{F}(\mathcal{D}, \sigma), k$
Output: $\mathcal{C}h(k, \mathcal{F}(\mathcal{D}, \sigma))$
1: $\mathcal{C}h(k, \mathcal{F}(\mathcal{D}, \sigma)) = \emptyset$
2: **for all** $\langle J, sup(J)\rangle \in \mathcal{F}(\mathcal{D}, \sigma)$ **do**
3: **for all** $I \subseteq J$ **do**
4: **compute** f_I^J;
5: **if** $0 < f_I^J < k$ **then**
6: **insert**$\langle \mathcal{C}_I^J, f_I^J\rangle$ **in** $\mathcal{C}h(k, \mathcal{F}(\mathcal{D}, \sigma))$;

Algorithm 1 detects all possible inference channels $\mathcal{C}h(k, \mathcal{F}(\mathcal{D}, \sigma))$ that hold in a collection of frequent itemsets $\mathcal{F}(\mathcal{D}, \sigma)$ by checking all possible pairs of itemsets $I, J \in \mathcal{F}(\mathcal{D}, \sigma)$ such that $I \subseteq J$. This could result in a very large number of checks. Suppose that $\mathcal{F}(\mathcal{D}, \sigma)$ is formed only by a maximal itemset Y and all its subsets (an itemset is maximal if none of its proper supersets is in $\mathcal{F}(\mathcal{D}, \sigma)$). If $|Y| = n$ we get $|\mathcal{F}(\mathcal{D}, \sigma)| = 2^n$ (we also count the empty set), while the number of possible $\mathcal{C}_I^J$ is $\sum_{1 \leq i \leq n} \binom{n}{i} (2^i - 1)$. In the following Section we study some interesting properties that allow to dramatically reduce the number of checks needed to retrieve $\mathcal{C}h(k, \mathcal{F}(\mathcal{D}, \sigma))$.

4 A Condensed Representation of Inference Channels

In this section we introduce a condensed representation of $\mathcal{C}h(k, \mathcal{F}(\mathcal{D}, \sigma))$, i.e., a subset of $\mathcal{C}h(k, \mathcal{F}(\mathcal{D}, \sigma))$ which is more efficient to compute, and sufficient to reconstruct the whole $\mathcal{C}h(k, \mathcal{F}(\mathcal{D}, \sigma))$. The benefits of having such condensed representation go far beyond mere efficiency. In fact, removing the redundancy existing in $\mathcal{C}h(k, \mathcal{F}(\mathcal{D}, \sigma))$, we also implicitly avoid redundant sanitization, when we will block inference channels holding in $\mathcal{F}(\mathcal{D}, \sigma)$ (recall that, as stated before, the issue of how to block inference channels is not covered in this paper).

Consider, for instance, the two inference channels $\langle \mathcal{C}_{ad}^{acd}, 1\rangle$ and $\langle \mathcal{C}_{abd}^{abcd}, 1\rangle$ holding in the database in Fig. 1(a): one is more specific than the other, but they both uniquely identify transaction t_7. It is easy to see that many other families of equivalent, and thus redundant, inference channels can be found. *How can we directly identify one and only one representative inference channel in each family of equivalent ones?* The theory of *closed itemsets* can help us with this problem.

Closed itemsets were first introduced in [12] and since then they have received a great deal of attention especially by an algorithmic point of view [17, 13]. They are a concise and lossless representation of all frequent itemsets, i.e., they contain the same information without redundancy. Intuitively, a closed itemset groups together all its subsets that have its same support; or in other words, it groups together itemsets which identify the same group of transactions.

Definition 13. Given the function $f(T) = \{i \in \mathcal{I} \mid \forall t \in T, i \in t\}$, which returns all the items included in the set of transactions T, and the function $g(X) = \{t \in \mathcal{T} \mid \forall i \in X, i \in t\}$ which returns the set of transactions supporting a given itemset X, the composite function $c = f \circ g$ is the *closure operator*. An itemset I is *closed* iff and only

if $c(I) = I$. Given a database $\mathcal{D}$ and a minimum support threshold σ, the set of *frequent closed itemsets* is denoted $\mathcal{C}l(\mathcal{D}, \sigma)$. An itemset $I \in \mathcal{C}l(\mathcal{D}, \sigma)$ is said to be *maximal* iff $\nexists J \supset I$ s.t. $J \in \mathcal{C}l(\mathcal{D}, \sigma)$.

Analogously to what happens for the pattern class of itemsets, if we consider the pattern class of conjunctive patterns we can rely on the *anti-monotonicity property of frequency*. For instance, the number of transactions for which the pattern $a \wedge \neg c$ holds is always larger than the number of transactions for which the pattern $a \wedge b \wedge \neg c \wedge \neg d$ holds.

Definition 14. Given two inference channels $\mathcal{C}_I^J$ and $\mathcal{C}_H^L$ we say that $\mathcal{C}_I^J \preceq \mathcal{C}_H^L$ when $I \subseteq H$ and $(J \setminus I) \subseteq (L \setminus H)$.

Proposition 15. $\mathcal{C}_I^J \preceq \mathcal{C}_H^L \Rightarrow \forall \mathcal{D}\ .\ f_I^J(\mathcal{D}) \geq f_H^L(\mathcal{D})$.

Therefore, when detecting inference channels, whenever we find a $\mathcal{C}_H^L$ such that $f_H^L(\mathcal{D}) \geq k$, we can avoid checking the support of all inference channels $\mathcal{C}_I^J \preceq \mathcal{C}_H^L$, since they will be k-anonymous.

Definition 16. An inference channel $\mathcal{C}_I^J$ is said to be *maximal* w.r.t. $\mathcal{D}$ and σ, if $\forall H, L$ such that $I \subseteq H$ and $(J \setminus I) \subseteq (L \setminus H)$, $f_H^L = 0$. The set of maximal inference channels is denoted $\mathcal{MC}h(k, \mathcal{C}l(\mathcal{D}, \sigma))$.

Proposition 17. $\mathcal{C}_I^J \in \mathcal{MC}h(k, \mathcal{C}l(\mathcal{D}, \sigma)) \Rightarrow I \in \mathcal{C}l(\mathcal{D}, \sigma) \wedge J$ *is maximal.*

Proof. $i)$ $I \in \mathcal{C}l(\mathcal{D}, \sigma)$: if I is not closed then consider its closure $c(I)$ and consider $J' = J \cup (c(I) \setminus I)$. For the definition of closure, the set of transactions containing I is the same of the set of transactions containing $c(I)$, and the set of transactions containing J' is the same of the set of transactions containing J. It follows that $\mathcal{C}_{c(I)}^{J'} \succeq \mathcal{C}_I^J$ and $f_{c(I)}^J = f_I^J > 0$. Then, if I is not closed, $\mathcal{C}_I^J$ is not maximal.

$ii)$ J is maximal: if J is not maximal then consider its frequent superset $J' = J \cup \{a\}$ and consider $I' = I \cup a$. It is straightforward to see that $f_I^J = f_I^{J'} + f_{I'}^{J'}$ and that $\mathcal{C}_I^{J'} \succeq \mathcal{C}_I^J$ and $\mathcal{C}_{I'}^{J'} \succeq \mathcal{C}_I^J$. Therefore, since $f_I^J > 0$, at least one among $f_I^{J'}$ and $f_{I'}^{J'}$ must be not null. Then, if J is not maximal, $\mathcal{C}_I^J$ is not maximal as well.

The next Theorem shows how the support of any channel in $\mathcal{C}h(k, \mathcal{F}(\mathcal{D}, \sigma))$ can be reconstructed from $\mathcal{MC}h(k, \mathcal{C}l(\mathcal{D}, \sigma))$.

Theorem 18. *Given $\mathcal{C}_I^J \in \mathcal{C}h(k, \mathcal{F}(\mathcal{D}, \sigma))$, let M be any maximal itemset such that $M \supseteq J$. The following equation holds:*

$$f_I^J(\mathcal{D}) = \sum_{c(X)} f_{c(X)}^M(\mathcal{D})$$

where $c(I) \subseteq c(X) \subseteq M$ and $c(X) \cap (J \setminus I) = \emptyset$.

Proof. See [5].

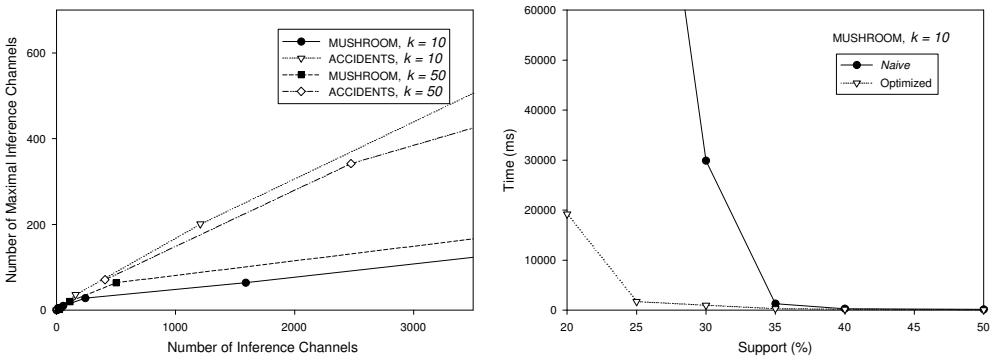
From Theorem 18 we conclude that all the addends needed to compute $f_I^J(\mathcal{D})$ for an inference channel are either in $\mathcal{MC}h(k, \mathcal{C}l(\mathcal{D}, \sigma))$ or are null. Therefore, as the set of all closed frequent itemsets $\mathcal{C}l(\mathcal{D}, \sigma)$ contains all the information of $\mathcal{F}(\mathcal{D}, \sigma)$ in a more compact representation, analogously the set $\mathcal{MC}h(k, \mathcal{C}l(\mathcal{D}, \sigma))$ represents, without redundancy, all the information in $\mathcal{C}h(k, \mathcal{F}(\mathcal{D}, \sigma))$.

In the database $\mathcal{D}$ of Figure 1(a), given $\sigma = 6$ and $k = 3$, $|\mathcal{C}h(3, \mathcal{F}(\mathcal{D}, 6))| = 58$ while $|\mathcal{MC}h(3, \mathcal{C}l(\mathcal{D}, 6))| = 5$ (Figure 1(e)), a reduction of one order of magnitude which is also confirmed by our experiments on real datasets, as reported in Figure 2(a). Moreover, in order to detect all inference channels holding in $\mathcal{F}(\mathcal{D}, \sigma)$, we can limit ourselves to retrieve only the inference channels in $\mathcal{MC}h(k, \mathcal{C}l(\mathcal{D}, \sigma))$, thus taking in input $\mathcal{C}l(\mathcal{D}, \sigma)$ instead of $\mathcal{F}(\mathcal{D}, \sigma)$ and thus performing a much smaller number of checks. Algorithm 2 exploits the anti-monotonicity of frequency (Prop. 15) and the property of maximal inference channels (Prop. 17) to compute $\mathcal{MC}h(k, \mathcal{C}l(\mathcal{D}, \sigma))$ from $\mathcal{C}l(\mathcal{D}, \sigma)$. Thanks to these two properties, Algorithm 2 is much faster, dramatically outperforming the naive inference channel detector (Algorithm 1), and scaling well even for very low support thresholds, as reported in Figure 2(b).

Algorithm 2 Optimized Inference Channel Detector

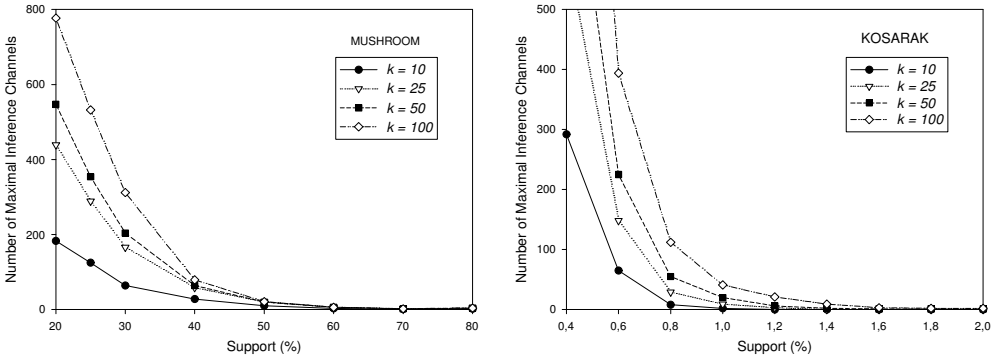
Input: $\mathcal{C}l(\mathcal{D}, \sigma), k$
Output: $\mathcal{MC}h(k, \mathcal{C}l(\mathcal{D}, \sigma))$

1: $M = \{I \in \mathcal{C}l(\mathcal{D}, \sigma) | I\ is\ maximal\}$;
2: $\mathcal{MC}h(k, \mathcal{C}l(\mathcal{D}, \sigma)) = \emptyset$;
3: **for all** $J \in M$ **do**
4: **for all** $I \in \mathcal{C}l(\mathcal{D}, \sigma)$ **such that** $I \subseteq J$ **do**
5: **compute** f_I^J;
6: **if** $0 < f_I^J < k$ **then**
7: **insert** $\langle \mathcal{C}_I^J, f_I^J \rangle$ **in** $\mathcal{MC}h(k, \mathcal{C}l(\mathcal{D}, \sigma))$;



(a) (b)

Fig. 2. Benefits of the condensed representation: size of the representations (a), and run time (b).



(a) (b)

Fig. 3. Experimental results on cardinality of $\mathcal{MC}h(k, \mathcal{C}l(\mathcal{D}, \sigma))$ on two datasets.

5 Anonymity vs. Accuracy: Empirical Observations

Algorithm 2 represents an optimized way to identify all threats to anonymity. Its performance revealed adequate in all our empirical evaluations using various datasets from the FIMI repository[2]; in all such cases the time improvement from the Naïve (Algorithm 1) to the optimized algorithm is about one order of magnitude. This level of efficiency allows an interactive-iterative use of the algorithm by the analyst, aimed at finding the best trade-off among privacy and accuracy of the collection of patterns. To be more precise, there is a conflict among keeping the support threshold as low as possible, in order to mine all interesting patterns, and avoiding the generation of anonymity threats. The best solution to this problem is precisely to find out the minimum support threshold that generates a collection of patterns with no threats. The plots in Figure 3 illustrate this point: on the x-axis we report the minimum support threshold, on the y-axis we report the total number of threats (the cardinality of $\mathcal{MC}h(k, \mathcal{C}l(\mathcal{D}, \sigma))$), and the various curves indicate such number according to different values of the anonymity threshold k. In Figure 3(a) we report the plot for the MUSHROOM dataset (a dense one), while in Figure 3(b) we report the plot for the KOSARAK dataset which is sparse. In both cases, it is evident the value of the minimum support threshold that represents the best trade-off, for any given value of k. However, in certain cases, the best support threshold can still be too high to mine a sufficient quantity of interesting patterns. In such cases, the only option is to allow lower support thresholds and then to block the inference channels in the mining outcome. This problem, as stated before, is not covered in this paper for lack of space, and will be presented in a forthcoming paper.

6 Conclusions

We introduced in this paper the notion of k-anonymous patterns. Such notion serves as a basis for a formal account of the intuition that a collection of patterns, obtained by data mining techniques and made available to the public, should not offer any possibilities to violate the privacy of the individuals whose data are stored in the source database. To

[2] `http://fimi.cs.helsinki.fi/data/`

the above aim, we formalized the threats to anonymity by means of inference channel through frequent itemsets, and provided practical algorithms to detect such channels.

Other issues, emerging from our approach, are worth a deeper investigation and are left to future research. These include: (i) a thorough comparison of the various different approaches that may be used to block inference channels; (ii) a comprehensive empirical evaluation of our approach: to this purpose we are conducting a large-scale experiment with real life bio-medical data about patients to assess both applicability and scalability of the approach in a realistic, challenging domain; (iii) an investigation whether the proposed notion of privacy-preserving pattern discovery may be generalized to other forms of patterns and models.

In any case, the importance of the advocated form of privacy-preserving pattern discovery is evident: demonstrably trustworthy data mining techniques may open up tremendous opportunities for new knowledge-based applications of public utility and large societal and economic impact.

References

1. D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the twentieth ACM PODS*, 2001.
2. R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD*.
3. R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD on Management of Data*.
4. R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the Twentieth VLDB*, 1994.
5. M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi. k-anonymous patterns. Technical Report 2005-TR-17, ISTI - C.N.R., 2005.
6. T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. In *Proceedings of the 6th PKDD*, 2002.
7. P. Fule and J. F. Roddick. Detecting privacy and ethical sensitivity in data mining results. In *Proc. of the 27th conference on Australasian computer science*, 2004.
8. D. Hand, H. Mannila, and P. Smyh. *Principles of Data Mining*. The MIT Press, 2001.
9. M. Kantarcioglu, J. Jin, and C. Clifton. When do data mining results violate privacy? In *Proceedings of the tenth ACM SIGKDD*, 2004.
10. D. Knuth. *Fundamental Algorithms*. Addison-Wesley, Reading, Massachusetts, 1997.
11. S. R. M. Oliveira, O. R. Zaiane, and Y. Saygin. Secure association rule sharing. In *Proc.of the 8th PAKDD*, 2004.
12. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *Proc. ICDT '99*, 1999.
13. J. Pei, J. Han, and J. Wang. Closet+: Searching for the best strategies for mining frequent closed itemsets. In *SIGKDD '03*, 2003.
14. L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty Fuzziness and Knowledge-based Systems*, 10(5), 2002.
15. L. Sweeney. k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty Fuzziness and Knowledge-based Systems*, 10(5), 2002.
16. V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *SIGMOD Rec.*, 33(1):50–57, 2004.
17. M. J. Zaki and C.-J. Hsiao. Charm: An efficient algorithm for closed itemsets mining. In *2nd SIAM International Conference on Data Mining*, 2002.