Efficient Discovery of Association Rules and Frequent Itemsets through Sampling with Tight Performance Guarantees

MATTEO RIONDATO and ELI UPFAL, Brown University

The tasks of extracting (top-K) Frequent Itemsets (FIs) and Association Rules (ARs) are fundamental primitives in data mining and database applications. Exact algorithms for these problems exist and are widely used, but their running time is hindered by the need of scanning the entire dataset, possibly multiple times. High-quality approximations of FIs and ARs are sufficient for most practical uses. Sampling techniques can be used for fast discovery of approximate solutions, but works exploring this technique did not provide satisfactory performance guarantees on the quality of the approximation due to the difficulty of bounding the probability of under- or oversampling any one of an unknown number of frequent itemsets. We circumvent this issue by applying the statistical concept of *Vapnik-Chervonenkis (VC) dimension* to develop a novel technique for providing tight bounds on the sample size that guarantees approximation of the (top-K) FIs and ARs within user-specified parameters. The resulting sample size is linearly dependent on the VC-dimension of a range space associated with the dataset. We analyze the VC-dimension of this range space and show that it is upper bounded by an easy-to-compute characteristic quantity of the dataset, the *d-index*, namely, the maximum integer d such that the dataset contains at least d transactions of length at least d such that no one of them is a superset of or equal to another. We show that this bound is tight for a large class of datasets. The resulting sample size is a significant improvement over previous known results. We present an extensive experimental evaluation of our technique on real and artificial datasets, demonstrating the practicality of our methods, and showing that they achieve even higher quality approximations than what is guaranteed by the analysis.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms: Algorithms, Theory, Performance, Experimentation

Additional Key Words and Phrases: Association rules, data mining, frequent itemsets, sampling, VC-dimension

ACM Reference Format:
Matteo Riondato and Eli Upfal. 2014. Efficient discovery of association rules and frequent itemsets through sampling with tight performance guarantees. ACM Trans. Knowl. Discov. Data 8, 4, Article 20 (July 2014), 32 pages.
DOI: http://dx.doi.org/10.1145/2629586

1. INTRODUCTION

Discovery of Frequent Itemsets (FIs) and Association Rules (ARs) is a fundamental computational primitive with application in data mining (market basket analysis),

A preliminary report of this work appeared in the proceedings of ECML PKDD 2012 as [Riondato and Upfal 2012].
This work is supported in part by the National Science Foundation, under awards IIS-0905553 and IIS-1247581.
Author's addresses: M. Riondato and E. Upfal, Department of Computer Science, Brown University, 115 Waterman Street, Providence, RI 02912, USA; email: {matteo, eli}@cs.brown.edu.

DOI: http://dx.doi.org/10.1145/2629586

databases (histogram construction), networking (heavy hitters), and more [Han et al. 2007, Sect. 5]. Depending on the particular application, one is interested in finding all itemsets with frequency greater than or equal to a user-defined threshold (FIs), identifying the K most FIs (top-K), or computing all ARs with user-defined minimum support and confidence level (see Sections 5.4 and 5.5 for additional criteria). Exact solutions to these problems require scanning the entire dataset, possibly multiple times. For large datasets that do not fit in main memory, this can be prohibitively expensive. Furthermore, such extensive computation is often unnecessary because high-quality approximations are sufficient for most practical applications. Indeed, a number of recent papers (see Section 2 for more details) explored the application of sampling for approximate solutions to these problems. However, the efficiency and practicality of the sampling approach depends on a tight relation between the size of the sample and the quality of the resulting approximation. Previous works do not provide satisfactory solutions to this problem.

The technical difficulty in analyzing any sampling technique for FI discovery problems is that, a priori, any subset of items can be among the most frequent ones, and the number of subsets is exponential in the number of distinct items appearing in the dataset. A standard analysis begins with a bound on the probability that a given itemset is either over- or underrepresented in the sample. Such a bound is easy to obtain using a large deviation bound such as the Chernoff bound or the Central Limit theorem [Mitzenmacher and Upfal 2005]. The difficulty is in combining the bounds for individual itemsets into a global bound that holds simultaneously for all the itemsets. A simple application of the union bound vastly overestimates the error probability because of the large number of possible itemsets, a large fraction of which may not be present in the dataset and therefore should not be considered. More sophisticated techniques, developed in recent works [Chakaravarthy et al. 2009; Pietracaprina et al. 2010; Chuang et al. 2005], give better bounds only in limited cases. A loose bound on the required sample size for achieving the user-defined performance guarantees decreases the gain obtained from the use of sampling.

In this work, we circumvent this problem through a novel application of the *Vapnik-Chervonenkis (VC)* dimension concept, a fundamental tool in statistical learning theory. Roughly speaking, the VC-dimension of a collection of indicator functions (a range space) is a measure of its complexity or expressiveness (see Section 3.2 for formal definitions). A major result [Vapnik and Chervonenkis 1971] relates the VC-dimension of a range space to a sufficient size for a random sample to simultaneously approximate all the indicator functions within predefined parameters. The main obstacle in applying the VC-dimension theory to particular computation problems is computing the VC-dimension of the range spaces associated with these problems.

We apply the VC-dimension theory to FI problems by viewing the presence of an itemset in a transaction as the outcome of an indicator function associated with the itemset. The major theoretical contributions of our work are a complete characterization of the VC-dimension of the range space associated with a dataset and a tight bound to this quantity. We prove that the VC-dimension is upper bounded by a characteristic quantity of the dataset that we call a *d-index*. The d-index is the maximum integer d such that the dataset contains at least d different transactions of length at least d such that no one of them is a subset of or equal to another in the considered set of transactions (see Definition 4.4). We show that this bound is tight by demonstrating a large class of datasets with a VC-dimension that matches the bound. Computing the d-index can be done in polynomial time, but it requires multiple scans of the dataset. We show how to compute an upper bound to the d-index with a single linear scan of the dataset in an online greedy fashion.

Table I. Required Sample Sizes (as Number of Transactions) for Various Approximations to FIs and ARs

Task/Approx.	This work	Best previous work
FIs/abs.	$\frac{4c}{\varepsilon^2}\left(v + \log\frac{1}{\delta}\right)$	$O\left(\frac{1}{\varepsilon^2}\left(\lvert\mathcal{I}\rvert + \log\frac{1}{\delta}\right)\right)^{\dagger}$
FIs/rel.	$\frac{4(2+\varepsilon)c}{\varepsilon^2(2-\varepsilon)\theta}\left(v\log\frac{2+\varepsilon}{\theta(2-\varepsilon)} + \log\frac{1}{\delta}\right)$	$\frac{24}{\varepsilon^2(1-\varepsilon)\theta}\left(\Delta + 5 + \log\frac{4}{(1-\varepsilon)\theta\delta}\right)^{\ddagger}$
top-K FIs/abs.	$\frac{16c}{\varepsilon^2}\left(v + \log\frac{1}{\delta}\right)$	$O\left(\frac{1}{\varepsilon^2}\left(\lvert\mathcal{I}\rvert + \log\frac{1}{\delta}\right)\right)^{\S}$
top-K FIs/rel.	$\frac{4(2+\varepsilon)c'}{\varepsilon^2(2-\varepsilon)\theta}\left(v\log\frac{2+\varepsilon}{\theta(2-\varepsilon)} + \log\frac{1}{\delta}\right)$	not available
ARs/abs.	$O\left(\frac{(1+\varepsilon)}{\varepsilon^2(1-\varepsilon)\theta}\left(v\log\frac{1+\varepsilon}{\theta(1-\varepsilon)} + \log\frac{1}{\delta}\right)\right)$	not available
ARs/rel.	$\frac{16c'(4+\varepsilon)}{\varepsilon^2(4-\varepsilon)\theta}\left(v\log\frac{4+\varepsilon}{\theta(4-\varepsilon)} + \log\frac{1}{\delta}\right)$	$\frac{48}{\varepsilon^2(1-\varepsilon)\theta}\left(\Delta + 5 + \log\frac{4}{(1-\varepsilon)\theta\delta}\right)^{\P}$

†[Toivonen 1996; Jia and Lu 2005; Li and Gopalan 2005; Zhang et al. 2003], ‡[Chakaravarthy et al. 2009], §[Scheffer and Wrobel 2002; Pietracaprina et al. 2010], ¶[Chakaravarthy et al. 2009].
The FIs and ARs are presented as functions of the VC-dimension v, the maximum transaction length Δ, the number of items $|\mathcal{I}|$, the accuracy ε, the failure probability δ, the minimum frequency θ, and the minimum confidence γ. Note that $v \leq \Delta \leq |\mathcal{I}|$ (but $v < |\mathcal{I}|$). The constants c and c' are absolute, with $c \leq 0.5$. See Section 5.4 for the sample sizes for approximations of the collection of ARs according to interestingness measures other than confidence.

The VC-dimension approach provides a unified tool for analyzing the various FI and AR problems (i.e., the market basket analysis tasks). We use it to prove tight bounds on the required sample size for extracting FIs with a minimum frequency threshold, for mining the top-K FIs, and for computing the collection of ARs with minimum frequency and "interestingness" thresholds, where the interestingness can be expressed in terms of confidence, leverage, lift, or other measure. Furthermore, we compute bounds for both absolute and relative approximations (see Section 3.1 for definitions), and our results extend to a variety of other measures proposed in the literature (see Section 5.4). We show that high-quality approximations can be obtained by mining a very small random sample of the dataset. Table I compares our technique to the best previously known results for the various problems (see Section 3.1 for definitions). Our bounds, which are linear in the VC-dimension associated with the dataset, are consistently smaller than previous results and less dependent on other parameters of the problem, such as the minimum frequency threshold and the dataset size. An extensive experimental evaluation demonstrates the advantage of our technique in practice.

This work is the first to provide a characterization and an explicit bound for the VC-dimension of the range space associated with a dataset and to apply the result to the extraction of FIs and ARs from a random sample of the dataset. We believe that this connection with statistical learning theory can be further exploited in other data mining problems.

Outline. We review relevant previous work in Section 2. In Section 3 we formally define the problem and our goals, and introduce definitions and lemmas used in the analysis. The main part of the analysis with derivation of a strict bound to the VC-dimension of ARs is presented in Section 4, while our algorithms and sample sizes for mining FIs, top-K FIs, and ARs through sampling are presented in Section 5. Section 6 contains an extensive experimental evaluation of our techniques. A discussion of our results and the conclusions can be found in Section 7.

2. RELATED WORK

Agrawal et al. [1993] introduced the problem of mining ARs in the basket data model, formalizing a fundamental task of information extraction in large datasets. Almost any

known algorithm for the problem starts by solving an FI problem and then generates the ARs implied by these FIs. Agrawal and Srikant [1994] presented *Apriori*, the most well-known algorithm for mining FIs, and *FastGenRules* for computing ARs from a set of itemsets. Various ideas for improving the efficiency of FI and AR algorithms have been studied, and we refer the reader to the survey by Ceglar and Roddick [2006] for a good presentation of recent contributions. However, the running times of all known algorithms heavily depend on the size of the dataset.

Mannila et al. [1994] were the first to propose the use of sampling to efficiently identify the collection of FIs, presenting some empirical results to validate the intuition. Toivonen [1996] presents an algorithm that, by mining a random sample of the dataset, builds a candidate set of FIs that contains all the FIs with a probability that depends on the sample size. There are no guarantees that all itemsets in the candidate set are frequent, but the set of candidates can be used to efficiently identify the set of FIs with at most two passes over the entire dataset. This work also suggests a bound on the sample size sufficient to ensure that the frequencies of itemsets in the sample are close to their real ones. The analysis uses Chernoff bounds and the union bound. The major drawback of this sample size is that it depends linearly on the number of individual items appearing in the dataset.

Zaki et al. [1997] show that static sampling is an efficient way to mine a dataset, but choosing the sample size using Chernoff bounds is too conservative, in the sense that it is possible to obtain the same accuracy and confidence in the approximate results at smaller sizes than what the theoretical analysis proves.

Other works tried to improve the bound to the sample size by using different techniques from statistics and probability theory, like the central limit theorem [Zhang et al. 2003; Li and Gopalan 2005; Jia and Lu 2005] or hybrid Chernoff bounds [Zhao et al. 2006].

Because theoretically derived bounds to the sample size were too loose to be useful, a corpus of works applied progressive sampling to extract FIs [John and Langley 1996; Chen et al. 2002; Parthasarathy 2002; Brönnimann et al. 2003; Chuang et al. 2005; Jia and Gao 2005; Wang et al. 2005a; Hwang and Kim 2006; Hu and Yu 2006; Mahafzah et al. 2009; Chen et al. 2011; Chandra and Bhaskar 2011]. Progressive sampling algorithms work by selecting a random sample and then trimming or enriching it by removing or adding new sampled transactions according to a heuristic or a self-similarity measure that is fast to evaluate until a suitable stopping condition is satisfied. The major downside of this approach is that it offers no guarantees on the quality of the obtained results.

Another approach to estimating the required sample size is presented by Chuang et al. [2008]. The authors give an algorithm that studies the distribution of frequencies of the itemsets and uses this information to fix a sample size for mining FIs, but without offering any theoretical guarantee.

A recent work by Chakaravarthy et al. [2009] gives the first analytical bound on a sample size that is linear in the length of the longest transaction, rather than in the number of items in the dataset. This work is also the first to present an algorithm that uses a random sample of the dataset to mine approximated solutions to the AR problem with quality guarantees. No experimental evaluation of their methods is presented, and they do not address the top-K FI problem. Our approach gives better bounds for the problems studied in Chakaravarthy et al. [2009] and applies to related problems such as the discovery of top-K FIs and absolute approximations.

Extracting the collection of top-K FIs is a more difficult task since the corresponding minimum frequency threshold is not known in advance [Cheung and Fu 2004; Fu et al. 2000]. Some works solved the problem by looking at *closed* top-K FIs, a concise

representation of the collection [Wang et al. 2005b; Pietracaprina and Vandin 2007], but they suffer from the same scalability problems as the algorithms for exactly mining FIs with a fixed minimum frequency threshold.

Previous works that used sampling to approximation the collection of top-K FIs [Scheffer and Wrobel 2002; Pietracaprina et al. 2010] used progressive sampling. Both works provide (similar) theoretical guarantees on the quality of the approximation. What is more interesting to us, both works present a theoretical upper bound to the sample size needed to compute such an approximation. The size depended linearly on the number of items. In contrast, our results give a sample size that only in the worst case is linear in the number of items but can be (and is, in practical cases) much less than that, depending on the dataset, a flexibility not provided by previous contributions. Sampling is used by Vasudevan and Vojonović [2009] to extract an approximation of the top-K frequent individual *items* from a sequence of items, which contains no item whose actual frequency is less than $f_K - \varepsilon$ for a fixed $0 < \varepsilon < 1$, where f_K is the *actual* frequency of the K-th most frequent item. They derive a sample size sufficient to achieve this result, but they assume the knowledge of f_K, which is rarely the case. An empirical sequential method can be used to estimate the right sample size. Moreover, the results cannot be directly extended to the mining of top-K frequent item(set)s from datasets of transactions with length greater than 1.

The *VC-dimension* was first introduced in a seminal article [Vapnik and Chervonenkis 1971] on the convergence of probability distributions, but it was only with the work of Haussler and Welzl [1986] and Blumer et al. [1989] that it was applied to the field of learning. Boucheron et al. [2005] present a good survey of the field with many recent advances. Since then, VC-dimension has encountered enormous success and application in the fields of computational geometry [Chazelle 2000; Matoušek 2002] and machine learning [Anthony and Bartlett 1999; Devroye et al. 1996]. Other applications include database management and graph algorithms. In the former, it was used in the context of constraint databases to compute good approximations of aggregate operators [Benedikt and Libkin 2002]. VC-dimension-related results were also recently applied in the field of database privacy by Blum et al. [2008] to show a bound on the number of queries needed for an attacker to learn a private concept in a database. Gross-Amblard [2011] showed that content with unbounded VC-dimension cannot be watermarked for privacy purposes. Riondato et al. [2011] computed an upper bound to the VC-dimension of classes of SQL queries and used it to develop a sampling-based algorithm for estimating the size of the output (selectivity) of queries run on a dataset. The results therein, although very different from that presented here due to the different settings, the different goals, and the different techniques used, inspired our present work. In the graph algorithms literature, VC-dimension has been used to develop algorithms to efficiently detect network failures [Kleinberg 2003; Kleinberg et al. 2008], balanced separators [Feige and Mahdian 2006], and events in a sensor networks [Gandhi et al. 2010], and to compute the shortest path [Abraham et al. 2011]. To our knowledge, this work is the first application of VC-dimension to knowledge discovery.

In this article, we extend our previous published work [Riondato and Upfal 2012] in a number of ways. The first prominent change is the development and analysis of a tighter bound to the VC-dimension of the range space associated with the dataset, together with a new polynomial time algorithm to compute such bounds and a very fast linear time algorithm to compute an upper bound. We present two novel methods to further speed up the computation of this quantity in Section 4.2. A new discussion about the relationship between these quantities can be found in Section 5.6. The second important change is the extension of our methods for approximating the collection of

ARs to measures of interestingness other than confidence (Section 5.4). In Section 5.5, we also discuss how effective our methods are in case one is interested in closed FIs. An interesting connection with the problem of monotone monomials is new and presented in Section 4.3. The proofs to most of our results were not published in the conference version but are presented here. We also added numerous examples to improve the understanding of the definitions and of the theoretical results, and we explain the connection of our results with other known results in statistical learning theory. As far as the experimental evaluation is concerned, we added comments on the precision and recall of our methods and on their scalability, which is also evident from their use inside a parallel/distributed algorithm for FI and AR mining [Riondato et al. 2012] for the MapReduce [Dean and Ghemawat 2004] platform that we describe in the conclusion.

3. PRELIMINARIES

This section introduces basic definitions and properties that will be used in later sections.

3.1. Datasets, Itemsets, and Association Rules

A *dataset* $\mathcal{D}$ is a collection of *transactions*, where each transaction τ is a subset of a ground set $\mathcal{I}$[1] There can be multiple identical transactions in $\mathcal{D}$. Elements of $\mathcal{I}$ are called *items,* and subsets of $\mathcal{I}$ are called *itemsets*. Let $|\tau|$ denote the number of items in transaction τ, which we call the *length* of τ. Given an itemset $A \subseteq \mathcal{I}$, the *support set* of A, denoted as $T_{\mathcal{D}}(A)$, is the set of transactions in $\mathcal{D}$ that contain A. The *support* of A, $s_{\mathcal{D}}(A) = |T_{\mathcal{D}}(A)|$, is the number of transaction in $\mathcal{D}$ that contains A, and the *frequency* of A, $f_{\mathcal{D}}(A) = |T_{\mathcal{D}}(A)|/|\mathcal{D}|$, is the fraction of transactions in $\mathcal{D}$ that contain A.

Definition 3.1. Given a *minimum frequency threshold* θ, $0 < \theta \leq 1$, the *FI's mining task with respect to* θ is finding all itemsets with frequency $\geq\theta$; that is, the set

$$\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta) = \{(A,\ f_{\mathcal{D}}(A)) : A \subseteq \mathcal{I} \text{ and } f_{\mathcal{D}}(A) \geq \theta\}.$$

To define the collection of top-K FIs, we assume a fixed *canonical ordering* of the itemsets in $2^{\mathcal{I}}$ by decreasing frequency in $\mathcal{D}$, with ties broken arbitrarily, and label the itemsets $A_1, A_2, \ldots, A_m$ according to this ordering. For a given $1 \leq K \leq m$, we denote by $f_{\mathcal{D}}^{(K)}$ the frequency $f_{\mathcal{D}}(A_K)$ of the K-th most FI A_K, and define the set of top-K FIs (with their respective frequencies) as

$$\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K) = \mathsf{FI}\big(\mathcal{D}, \mathcal{I}, f_{\mathcal{D}}^{(K)}\big).$$

One of the main uses of FIs is in the discovery of ARs. An *association rule* W is an expression "$A \Rightarrow B$" where A and B are itemsets such that $A \cap B = \emptyset$. The *support* $s_{\mathcal{D}}(W)$ (resp. frequency $f_{\mathcal{D}}(W)$) of the association rule W is the support (resp. frequency) of the itemset $A \cup B$. The *confidence* $c_{\mathcal{D}}(W)$ of W is the ratio $f_{\mathcal{D}}(A \cup B)/f_{\mathcal{D}}(A)$. Intuitively, an association rule "$A \Rightarrow B$" expresses, through its support and confidence, how likely it is for the itemset B to appear in the same transactions as itemset A. The confidence of the AR can be interpreted as the conditional probability of B being present in a transaction that contains A. Many other measures can be used to quantify the interestingness of an AR [Tan et al. 2004] (see also Section 5.4).

Definition 3.2. Given a dataset $\mathcal{D}$ with transactions built on a ground set $\mathcal{I}$, and given a minimum frequency threshold θ and a minimum confidence threshold γ, the

[1]We assume $\mathcal{I} = \cup_{\tau \in \mathcal{D}} \tau$, i.e., all the elements of $\mathcal{I}$ appear in at least one transaction from $\mathcal{D}$.

AR's task with respect to θ *and* γ is to identify the set

$$\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma) = \left\{(W,\, f_{\mathcal{D}}(W), c_{\mathcal{D}}(W))|\ \text{association rule } W,\, f_{\mathcal{D}}(W) \geq \theta, c_{\mathcal{D}}(W) \geq \gamma\right\}.$$

We say that an itemset A (resp. an association rule W) is in $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$ or in $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K)$ (resp. in $\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$) when there A (resp. W) is part of a pair in $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$ or $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K)$, (resp. a triplet $\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$).

In this work, we are interested in extracting absolute and relative approximations of the sets $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$, $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K)$, and $\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$.

Definition 3.3. Given a parameter ε_{abs} (resp. ε_{rel}), an *absolute* ε_{abs}*-close approximation* (resp. a *relative* ε_{rel}*-close approximation*) of $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$ is a set $\mathcal{C} = \{(A,\, f_A) : A \subseteq \mathcal{I},\, f_A \in [0, 1]\}$ of pairs $(A,\, f_A)$ where f_A approximates $f_{\mathcal{D}}(A)$. $\mathcal{C}$ is such that:

(1) $\mathcal{C}$ contains all itemsets appearing in $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$;
(2) $\mathcal{C}$ contains no itemset A with frequency $f_{\mathcal{D}}(A) < \theta - \varepsilon_{\text{abs}}$ (resp. $f_{\mathcal{D}}(A) < (1 - \varepsilon_{\text{rel}})\theta$);
(3) For every pair $(A,\, f_A) \in \mathcal{C}$, it holds that $|f_{\mathcal{D}}(A) - f_A| \leq \varepsilon_{\text{abs}}$ (resp. $|f_{\mathcal{D}}(A) - f_A| \leq \varepsilon_{\text{rel}} f_{\mathcal{D}}(A)$).

As an example, consider a dataset $\mathcal{D}$ in which transactions have all length 1 and are built on the ground set $\mathcal{I} = \{a, b, c, d\}$. Suppose that $f_{\mathcal{D}}(a) = 0.4$, $f_{\mathcal{D}}(b) = 0.3$, $f_{\mathcal{D}}(c) = 0.2$, and $f_{\mathcal{D}}(d) = 0.1$ (clearly there are no other itemsets). If we set $\theta = 0.22$ and $\varepsilon = 0.05$, an absolute ε-close approximation C of $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$ *must* contain two pairs $(a,\, f_a)$ and $(b,\, f_b)$ as $a, b \in \mathsf{FI}(Ds, \mathcal{I}, \theta)$. At the same time, C *might* contain a pair $(c,\, f_c)$ because $f_{\mathcal{D}}(c) > \theta - \varepsilon$. On the other hand, C *must not* contain a pair $(d,\, f_d)$ because $f_{\mathcal{D}}(d) < \theta - \varepsilon$. The values f_a, f_b, and eventually f_c must be not more than ε far from $f_{\mathcal{D}}(a)$, $f_{\mathcal{D}}(b)$, and $f_{\mathcal{D}}(c)$, respectively.

This definition extends easily to the case of top-K FI mining using the equivalence

$$\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K) = \mathsf{FI}\left(\mathcal{D}, \mathcal{I}, f_{\mathcal{D}}^{(K)}\right):$$

an absolute (resp. relative) ε-close approximation to $\mathsf{FI}(\mathcal{D}, \mathcal{I}, f_{\mathcal{D}}^{(K)})$ is an absolute (resp. relative) ε-close approximation to $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K)$.

For the case of ARs, we have the following definition.

Definition 3.4. Given a parameter ε_{abs} (resp. ε_{rel}), an *absolute* ε_{abs}*-close approximation* (resp. a *relative* ε_{rel}*-close approximation*) of $\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ is a set

$$\mathcal{C} = \{(W,\, f_W, c_W)\text{: association rule } W,\, f_W \in [0, 1], c_W \in [0, 1]\}$$

of triplets $(W,\, f_W, c_W)$, where f_W and c_W approximate $f_{\mathcal{D}}(W)$ and $c_{\mathcal{D}}(W)$, respectively. $\mathcal{C}$ is such that:

(1) $\mathcal{C}$ contains all association rules appearing in $\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$;
(2) $\mathcal{C}$ contains no association rule W with frequency $f_{\mathcal{D}}(W) < \theta - \varepsilon_{\text{abs}}$ (resp. $f_{\mathcal{D}}(W) < (1 - \varepsilon_{\text{rel}})\theta$);
(3) For every triplet $(W,\, f_W, c_W) \in \mathcal{C}$, it holds that $|f_{\mathcal{D}}(W) - f_W| \leq \varepsilon_{\text{abs}}$ (resp. $|f_{\mathcal{D}}(W) - f_W| \leq \varepsilon_{\text{rel}}\theta$).
(4) $\mathcal{C}$ contains no association rule W with confidence $c_{\mathcal{D}}(W) < \gamma - \varepsilon_{\text{abs}}$ (resp. $c_{\mathcal{D}}(W) < (1 - \varepsilon_{\text{rel}})\gamma$);
(5) For every triplet $(W,\, f_W, c_W) \in \mathcal{C}$, it holds that $|c_{\mathcal{D}}(W) - c_W| \leq \varepsilon_{\text{abs}}$ (resp. $|c_{\mathcal{D}}(W) - c_W| \leq \varepsilon_{\text{rel}} c_{\mathcal{D}}(W)$).

Note that the definition of relative ε-close approximation to $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$ (resp. to $\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$) is more stringent than the definition of ε-close solution to FI mining (resp. association rule mining) in Chakaravarthy et al. [2009, Sect. 3]. Specifically, we

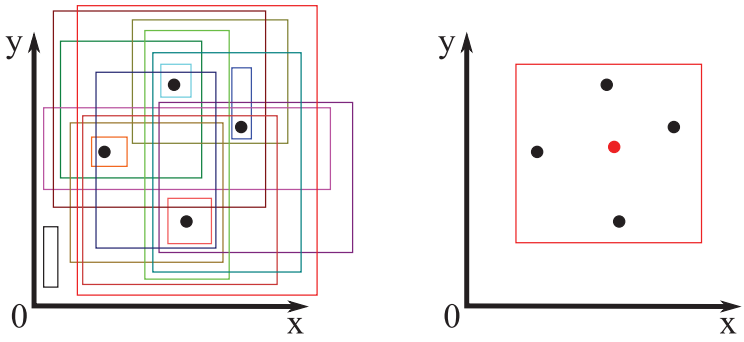


Fig. 1. Example of range space and VC-dimension. The space of points is the plane $\mathbb{R}^2$, and the set of ranges is the set of all *axis-aligned rectangles*. The figure on the left shows graphically that it is possible to shatter a set of four points using 16 rectangles. On the right, instead, one can see that it is impossible to shatter five points because, for any choice of the five points, there will always be one (the red point in the figure) that is internal to the convex hull of the other four, so it would be impossible to find an axis-aligned rectangle containing the four points but not the internal one. Hence $\mathsf{VC}((X, R)) = 4$.

require an approximation of the frequencies (and confidences) in addition to the approximation of the collection of itemsets or association rules (Property 3 in Definition 3.3 and Properties 3 and 5 in Definition 3.4).

3.2. VC-Dimension

The VC-dimension of a space of points is a measure of the complexity or expressiveness of a family of indicator functions (or equivalently a family of subsets) defined on that space [Vapnik and Chervonenkis 1971]. A finite bound on the VC-dimension of a structure implies a bound on the number of random samples required for approximately learning that structure. We outline here some basic definitions and results and refer the reader to the works of Alon and Spencer [2008, Sect. 14.4], Devroye et al. [1996], and Vapnik [1999] for more details on VC-dimension. See Section 2 for applications of VC-dimension in computer science.

We define a *range space* as a pair (X, R) where X is a (finite or infinite) set and R is a (finite or infinite) family of subsets of X. The members of X are called *points,* and those of R are called *ranges*. Given $A \subset X$, the *projection* of R on A is defined as $P_R(A) = \{r \cap A : r \in R\}$. If $P_R(A) = 2^A$, then A is said to be *shattered by* R. The VC-dimension of a range space is the cardinality of the largest set shattered by the space.

Definition 3.5. Let $S = (X, R)$ be a range space. The V-C dimension of S, denoted as $\mathsf{VC}(S)$, is the maximum cardinality of a shattered subset of X. If there are arbitrary large shattered subsets, then $\mathsf{VC}(S) = \infty$.

Note that a range space (X, R) with an arbitrary large set of points X and an arbitrary large family of ranges R can have a bounded VC-dimension. A simple example is the family of intervals in $[0, 1]$ (i.e., X is all the points in $[0, 1]$ and R all the intervals $[a, b]$, such that $0 \leq a \leq b \leq 1$). Let $A = \{x, y, z\}$ be the set of three points $0 < x < y < z < 1$. No interval in R can define the subset $\{x, z\}$ so the VC-dimension of this range space is less than 3 [Matoušek 2002, Lemma 10.3.1]. Another example is shown in Figure 1.

The main application of VC-dimension in statistics and learning theory is its relation to the size of the sample needed to approximate learning the ranges, in the following sense.

Definition 3.6. Let (X, R) be a range space and let A be a finite subset of X. For $0 < \varepsilon < 1$, a subset $B \subset A$ is an *ε-approximation* for A if for all $r \in R$, we have

$$\left| \frac{|A \cap r|}{|A|} - \frac{|B \cap r|}{|B|} \right| \leq \varepsilon. \tag{1}$$

A similar definition offers relative guarantees.

Definition 3.7. Let (X, R) be a range space and let A be a finite subset of X. For $0 < p, \varepsilon < 1$, a subset $B \subset A$ is a *relative (p, ε)-approximation* for A if for any range $r \in R$ such that $|A \cap r|/|A| \geq p$ we have

$$\left| \frac{|A \cap r|}{|A|} - \frac{|B \cap r|}{|B|} \right| \leq \varepsilon \frac{|A \cap r|}{|A|}$$

and for any range $r \in R$ such that $|A \cap r|/|A| < p$ we have $|B \cap r|/|B| \leq (1 + \varepsilon)p$.

An ε-approximation (resp. a relative (p, ε)-approximation) can be constructed by random sampling points of the point space [Har-Peled and Sharir 2011, Theorem 2.12 (resp. 2.11), see also [Li et al. 2001]].

THEOREM 3.8. *There is an absolute positive constant c (resp. c') such that if (X, R) is a range-space of VC-dimension at most v, $A \subset X$ is a finite subset, and $0 < \varepsilon, \delta < 1$ (resp. and $0 < p < 1$), then a random subset $B \subset A$ of cardinality m, where*

$$m \geq \min \left\{ |A|, \frac{c}{\varepsilon^2} \left(v + \log \frac{1}{\delta} \right) \right\}, \tag{2}$$

(resp. $m \geq \min\{|A|, c'\varepsilon^{-2}p^{-1}(v \log 1/p - \log 1/\delta)\}$) is an ε-approximation (resp. a relative (p, ε)-approximation) for A with probability at least $1 - \delta$.

Note that throughout the work we assume the sample to be drawn *with* replacement if $m < |A|$ (otherwise, the sample is exactly the set A). The constants c and c' are absolute and do not depend on the range space or on any other parameter. Löffler and Phillips [2009] estimated experimentally that the absolute constant c is at most 0.5. No upper bound is currently known for c'. Up to a constant, the bounds presented in Theorem 3.8 are tight [Li et al. 2001, Theorem 5].

It is also interesting to note that an ε-approximation of size $O(v\varepsilon^{-2}(\log v - \log \varepsilon))$ can be built *deterministically* in time $O(v^{3v}(\varepsilon^{-2}(\log v - \log \varepsilon))^v |X|)$ [Chazelle 2000].

4. THE DATASET'S RANGE SPACE AND ITS VC-DIMENSION

Our next step is to define a range space of the dataset and the itemsets. We will use this space together with Theorem 3.8 to compute the bounds to sample sizes sufficient to compute approximate solutions for the various tasks of market basket analysis.

Definition 4.1. Let $\mathcal{D}$ be a dataset of transactions that are subsets of a ground set $\mathcal{I}$. We define $S = (X, R)$ to be a range space associated with $\mathcal{D}$ such that:

(1) $X = \mathcal{D}$ is the set of transactions in the dataset.
(2) $R = \{T_{\mathcal{D}}(A) \mid A \subseteq \mathcal{I}, A \neq \emptyset\}$ is a family of sets of transactions such that for each nonempty itemset $A \subseteq \mathcal{I}$, the set $T_{\mathcal{D}}(A) = \{\tau \in \mathcal{D} | A \subseteq \tau\}$ of all transactions containing A is an element of R.

It is easy to see that in practice the collection R of ranges contains all and only the sets $T_{\mathcal{D}}(A)$ where A is a *closed itemset*; that is, a set such that for each nonempty $B \subseteq A$ we have $T_{\mathcal{D}}(B) = T_{\mathcal{D}}(A)$ and for any $C \supset A$, $T_{\mathcal{D}}(C) \subsetneq T_{\mathcal{D}}(A)$. Closed itemsets are used to summarize the collection of FIs [Calders et al. 2006].

The VC-dimension of this range space is the maximum size of a set of transactions that can be shattered by the support sets of the itemsets, as expressed by the following theorem and the following corollary.

THEOREM 4.2. *Let $\mathcal{D}$ be a dataset and let $S = (X, R)$ be the associated range space. Let $v \in \mathbb{N}$. Then, $\mathsf{VC}(S) \geq v$ if and only if there exists a set $\mathcal{A} \subseteq \mathcal{D}$ of v transactions from $\mathcal{D}$ such that, for each subset $\mathcal{B} \subseteq \mathcal{A}$, there exists an itemset $I_{\mathcal{B}}$ such that the support set of $I_{\mathcal{B}}$ in $\mathcal{A}$ is exactly $\mathcal{B}$; that is, $T_{\mathcal{A}}(I_{\mathcal{B}}) = \mathcal{B}$.*

PROOF. "$\Leftarrow$". From the definition of $I_{\mathcal{B}}$, we have that $T_{\mathcal{D}}(I_{\mathcal{B}}) \cap \mathcal{A} = \mathcal{B}$. By definition of $P_R(\mathcal{A})$, this means that $\mathcal{B} \in P_R(\mathcal{A})$ for any subset $\mathcal{B}$ of $\mathcal{A}$. Then, $P_R(\mathcal{A}) = 2^{\mathcal{A}}$, which implies $\mathsf{VC}(S) \geq v$.

"$\Rightarrow$". Let $\mathsf{VC}(S) \geq v$. Then, by the definition of VC-dimension, there is a set $\mathcal{A} \subseteq \mathcal{D}$ of v transactions from $\mathcal{D}$ such that $P_R(A) = 2^{\mathcal{A}}$. By definition of $P_R(\mathcal{A})$, this means that for each subset $\mathcal{B} \subseteq \mathcal{A}$ there exists an itemset $I_{\mathcal{B}}$ such that $T_{\mathcal{D}}(I_{\mathcal{B}}) \cap \mathcal{A} = \mathcal{B}$. We want to show that no transaction $\rho \in \mathcal{A} \setminus \mathcal{B}$ contains $I_{\mathcal{B}}$. Assume now by contradiction that there is a transaction $\rho^* \in \mathcal{A} \setminus \mathcal{B}$ containing $I_{\mathcal{B}}$. Then, $\rho^* \in T_{\mathcal{D}}(I_{\mathcal{B}})$ and, given that $\rho^* \in \mathcal{A}$, we have $\rho^* \in T_{\mathcal{D}}(I_{\mathcal{B}}) \cap \mathcal{A}$. But by construction, we have that $T_{\mathcal{D}}(I_{\mathcal{B}}) \cap \mathcal{A} = \mathcal{B}$ and $\rho^* \notin \mathcal{B}$ because $\rho^* \in \mathcal{A} \setminus \mathcal{B}$. Then we have a contradiction, and there cannot be such a transaction ρ^*. □

COROLLARY 4.3. *Let $\mathcal{D}$ be a dataset and $S = (\mathcal{D}, R)$ be the corresponding range space. Then the VC-dimension $\mathsf{VC}(S)$ of S is the maximum integer v such that there is a set $\mathcal{A} \subseteq \mathcal{D}$ of v transactions from $\mathcal{D}$, such that for each subset $\mathcal{B} \subseteq \mathcal{A}$ of $\mathcal{A}$, there exists an itemset $I_{\mathcal{B}}$ such that the support of $I_{\mathcal{B}}$ in $\mathcal{A}$ is exactly $\mathcal{B}$; that is, $T_{\mathcal{A}}(I_{\mathcal{B}}) = \mathcal{B}$.*

For example, consider the dataset $\mathcal{D} = \{\{a, b, c, d\}, \{a, b\}, \{a, c\}, \{d\}\}$ of four transactions built on the set of items $\mathcal{I} = \{a, b, c, d\}$. It is easy to see that the set of transactions $\mathcal{A} = \{\{a, b\}, \{a, c\}\}$ can be shattered: $\mathcal{A} = \mathcal{A} \cap T_{\mathcal{D}}(\{a\})$, $\{\{a, b\}\} = \mathcal{A} \cap T_{\mathcal{D}}(\{a, b\})$, $\{\{a, c\}\} = \mathcal{A} \cap T_{\mathcal{D}}(\{a, c\})$, $\emptyset = \mathcal{A} \cap T_{\mathcal{D}}(\{d\})$. It should be clear that there is no set of three transactions in $\mathcal{D}$ that can be shattered, so the VC-dimension of the range space associated to $\mathcal{D}$ is exactly 2.

Computing the exact VC-dimension of the range space associated to a dataset is extremely expensive from a computational point of view. This does not come as a surprise because it is known that computing the VC-dimension of a range space (X, R) can take time $O(|R||X|^{\log |R|})$ [Linial et al. 1991, Theorem 4.1]. It is instead possible to give an upper bound to the VC-dimension and a procedure to efficiently compute the bound.

We now define a characteristic quantity of the dataset, called the *d-index,* and show that it is a tight bound to the VC-dimension of the range space associated to the dataset, then present an algorithm to efficiently compute an upper bound to the d-index with a single linear scan of the dataset.

Definition 4.4. Let $\mathcal{D}$ be a dataset. The *d-index* of $\mathcal{D}$ is the maximum integer d such that $\mathcal{D}$ contains at least d different transactions of length at least d such that no one of them is a subset of another; that is, the transactions form an *antichain*.

Consider now the dataset $\mathcal{D} = \{\{a, b, c, d\}, \{a, b, d\}, \{a, c\}, \{d\}\}$ of four transactions built on the set of items $\mathcal{I} = \{a, b, c, d\}$. The d-index of $\mathcal{D}$ is 2, as the transactions $\{a, b, d\}$ and $\{a, c\}$ form an antichain. Note that the antichain determining the d-index is not necessarily the largest antichain that can be built on the transactions of $\mathcal{D}$. For example, if $\mathcal{D} = \{\{a, b, c, d\}, \{a, b\}, \{a, c\}, \{a\}, \{b\}, \{c\}, \{d\}\}$, the largest antichain would be $\{\{a\}, \{b\}, \{c\}, \{d\}\}$, but the antichain determining the d-index of the dataset would be $\{\{a, b\}, \{a, c\}, \{d\}\}$.

Intuitively, the reason for considering an antichain of transactions is that, if τ is a transaction that is a subset of another transaction τ', ranges containing τ' necessarily also contain τ (the opposite is not necessarily true), so it would be impossible to shatter a set containing both transactions.

It is easy to see that the d-index of a dataset built on a set of items $\mathcal{I}$ is at most equal to the length of the longest transaction in the dataset and, in any case, no greater than $|\mathcal{I}| - 1$.

The d-index is an upper bound to the VC-dimension of a dataset.

THEOREM 4.5. *Let $\mathcal{D}$ be a dataset with d-index d. Then, the range space $S = (X, R)$ corresponding to $\mathcal{D}$ has VC-dimension at most d.*

PROOF. Let $\ell > d$ and assume that S has VC-dimension ℓ. From Definition 3.5, there is a set $\mathcal{K}$ of ℓ transactions of $\mathcal{D}$ that is shattered by R. Clearly, $\mathcal{K}$ cannot contain any transaction equal to $\mathcal{I}$ because such a transaction would appear in all ranges of R, and so it would not be possible to shatter $\mathcal{K}$. At the same time, for any two transactions τ, τ' in $\mathcal{K}$, we must have neither $\tau \subseteq \tau'$ nor $\tau' \subseteq \tau$, otherwise the shorter transaction of the two would appear in all ranges where the longer one appears, and so it would not be possible to shatter $\mathcal{K}$. Then, $\mathcal{K}$ must be an antichain. From this and from the definitions of d and ℓ, $\mathcal{K}$ must contain a transaction τ such that $|\tau| \leq d$. The transaction τ is a member of $2^{\ell-1}$ subsets of $\mathcal{K}$. We denote these subsets of $\mathcal{K}$ containing τ as $\mathcal{A}_i$, $1 \leq i \leq 2^{\ell-1}$, labeling them in an arbitrary order. Since $\mathcal{K}$ is shattered (i.e., $P_R(\mathcal{K}) = 2^{\mathcal{K}}$), we have

$$\mathcal{A}_i \in P_R(\mathcal{K}), 1 \leq i \leq 2^{\ell-1}.$$

From this and the definition of $P_R(\mathcal{K})$, it follows that for each set of transactions $\mathcal{A}_i$ there must be a nonempty itemset B_i such that

$$T_{\mathcal{D}}(B_i) \cap \mathcal{K} = \mathcal{A}_i \in P_R(\mathcal{K}). \tag{3}$$

Because the $\mathcal{A}_i$ are all different from each other, this means that the $T_{\mathcal{D}}(B_i)$ are all different from each other, which in turn requires that the B_i be all different from each other, for $1 \leq i \leq 2^{\ell-1}$.

Since $\tau \in \mathcal{A}_i$ and $\tau \in \mathcal{K}$ by construction, it follows from Equation (3) that

$$\tau \in T_{\mathcal{D}}(B_i), 1 \leq i \leq 2^{\ell-1}.$$

From this and the definition of $T_{\mathcal{D}}(B_i)$, we get that all the itemsets $B_i, 1 \leq i \leq 2^{\ell-1}$ appear in the transaction τ. But $|\tau| \leq d < \ell$; therefore, τ can only contain at most $2^d - 1 < 2^{\ell-1}$ nonempty itemsets, whereas there are $2^{\ell-1}$ different itemsets B_i.

This is a contradiction; therefore, our assumption is false, and $\mathcal{K}$ cannot be shattered by R, which implies that $\mathsf{VC}(S) \leq d$. □

This bound is strict; that is, there are indeed datasets with VC-dimension exactly d, as formalized by the following theorem.

THEOREM 4.6. *There exists a dataset $\mathcal{D}$ with d-index d, and the corresponding range space has VC-dimension exactly d.*

PROOF. For $d = 1$, $\mathcal{D}$ can be any dataset with at least two different transactions $\tau = \{a\}$ and $\tau' = \{b\}$ of length 1. The set $\{\tau\} \subseteq \mathcal{D}$ is shattered because $T_{\mathcal{D}}(\{a\}) \cap \{\tau\} = \{\tau\}$ and $T_{\mathcal{D}}(\{b\}) \cap \{\tau\} = \emptyset$.

Without loss of generality, let the ground set $\mathcal{I}$ be $\mathbb{N}$. For a fixed $d > 1$, let $\tau_i = \{0, 1, 2, \ldots, i-1, i+1, \ldots, d\} 1 \leq i \leq d$ and consider the set of d transactions $\mathcal{K} = \{\tau_i, 1 \leq i \leq d\}$. Note that $|\tau_i| = d$ and $|\mathcal{K}| = d$ and there is no pair of transactions τ_i, τ_j with $i \neq j$ for which we have either $\tau_i \subseteq \tau_j$ or $\tau_j \subseteq \tau_i$.

$\mathcal{D}$ is a dataset containing $\mathcal{K}$ and any number of arbitrary transactions from $2^{\mathcal{I}}$ of length at most d. Let $S = (X, R)$ be the range space corresponding to $\mathcal{D}$. We now show that $\mathcal{K} \subseteq X$ is shattered by ranges from R, which implies $\mathsf{VC}(S) \geq d$.

For each $\mathcal{A} \in 2^{\mathcal{K}} \setminus \{\mathcal{K}, \emptyset\}$, let $Y_{\mathcal{A}}$ be the itemset

$$Y_{\mathcal{A}} = \{1, \ldots, d\} \setminus \{i : \tau_i \in \mathcal{A}\}.$$

Let $Y_{\mathcal{K}} = \{0\}$ and let $Y_{\emptyset} = \{d + 1\}$. By construction we have

$$T_{\mathcal{K}}(Y_{\mathcal{A}}) = \mathcal{A}, \forall \mathcal{A} \subseteq \mathcal{K};$$

that is, the itemset $Y_{\mathcal{A}}$ appears in all transactions in $\mathcal{A} \subseteq \mathcal{K}$ but not in any transaction from $\mathcal{K} \setminus \mathcal{A}$, for all $\mathcal{A} \in 2^{\mathcal{K}}$. This means that

$$T_{\mathcal{D}}(Y_{\mathcal{A}}) \cap \mathcal{K} = T_{\mathcal{K}}(Y_{\mathcal{A}}) = \mathcal{A}, \forall \mathcal{A} \subseteq \mathcal{K}.$$

Since for all $\mathcal{A} \subseteq \mathcal{K}$, $T_{\mathcal{D}}(Y_{\mathcal{A}}) \in R$ by construction, the equation just given implies that

$$\mathcal{A} \in P_R(\mathcal{K}), \forall \mathcal{A} \subseteq \mathcal{K}.$$

This means that $\mathcal{K}$ is shattered by R; hence, $\mathsf{VC}(S) \geq d$. From this and Theorem 4.5, we can conclude that $\mathsf{VC}(S) = d$. □

Consider again the dataset $\mathcal{D} = \{\{a, b, c, d\}, \{a, b\}, \{a, c\}, \{d\}\}$ of four transactions built on the set of items $\mathcal{I} = \{a, b, c, d\}$. We argued before that the VC-dimension of the range space associated to this dataset is exactly 2, and it is easy to see that the d-index of $\mathcal{D}$ is also 2.

4.1. Computing the d-Index of a Dataset

The d-index of a dataset $\mathcal{D}$ can be obtained exactly in polynomial time by computing, for each length ℓ, the size w_ℓ of the largest antichain that can be built using the transactions of length at least ℓ from $\mathcal{D}$. If $w \geq \ell$, then the d-index is at least ℓ. The maximum ℓ for which $w_\ell \geq \ell$ is the d-index of $\mathcal{D}$. The size of the largest antichain that can be built on the elements of a set can be computed by solving a maximum matching problem on a bipartite graph that has two nodes for each element of the set [Ford and Fulkerson 1962]. Computing the maximum matching can be done in polynomial time [Hopcroft and Karp 1973].

In practice, this approach can be quite slow because it requires, for each value taken by ℓ, a scan of the dataset to create the set of transactions of length at least ℓ and to solve a maximum matching problem. Hence, we now present an algorithm to efficiently compute an upper bound q to the d-index with a single linear scan of the dataset and with $O(q)$ memory.

It is easy to see that the d-index of a dataset $\mathcal{D}$ is upper bounded by the maximum integer q such that $\mathcal{D}$ contains at least q different (i.e., not containing the same items) transactions of length at least q and less than $|\mathcal{I}|$. This upper bound, which we call *d-bound*, ignores the constraint that the transactions that concur to the computation of the d-index must form an antichain. We can compute the d-bound in a greedy fashion by scanning the dataset once and keeping in memory the maximum integer q such that we saw at least q transactions of length q until this point of the scanning. We also keep in memory the q longest different transactions to avoid counting transactions that are equal to ones we have already seen because, as we already argued, a set containing identical transactions cannot be shattered. Copies of a transaction should not be included in the computation of the d-index, so it is not useful to include them in the computation of the d-bound. The pseudocode for computing the d-bound in the way we just described is presented in Algorithm 1. The function `getNextTransaction` returns one transaction at a time from the dataset. Note, though, that this does not imply that,

ALGORITHM 1: Compute the d-bound, an upper bound to the d-index of adataset

Input: a dataset $\mathcal{D}$
Output: the d-bound q, an upper bound to the d-index of $\mathcal{D}$

```
1  τ ← getNextTransaction(D)
2  T ← {τ}
3  q ← 1
4  while scanIsNotComplete() do
5  |   τ ← getNextTransaction(D)
6  |   if |τ| > q and τ ≠ I and ¬∃a ∈ T such that τ = a then
7  |   |   R ← T ∪ {τ}
8  |   |   q ← max integer such that R contains at least q transactions of length at least q
9  |   |   T ← set of the q longest transactions from R (ties broken arbitrarily)
10 |   end
11 end
12 return q
```

in a disk-based system, the algorithm needs a random read for each transaction. If the dataset is stored in a block-based fashion, one can read one block at a time and scan all transactions in that block, given that the order in which the transactions are scanned is not relevant for the correctness of the algorithm. Thus, in the worst case, the algorithm performs a random read per block. The following lemma deals with the correctness of the algorithm.

LEMMA 4.7. *The algorithm presented in Algorithm 1 computes the maximum integer q such that $\mathcal{D}$ contains at least q different transactions of length at least q and less than $|\mathcal{I}|$.*

PROOF. The algorithm maintains the following invariant after each update of $\mathcal{T}$: The set $\mathcal{T}$ contains the ℓ longest (ties broken arbitrarily) different transactions of length at least ℓ, where ℓ is the maximum integer r for which, up to this point of the scan, the algorithm saw at least r different transactions of length at least r. It should be clear that if the invariant holds after the scanning is completed, the thesis follows because the return value q is exactly the size $|\mathcal{T}| = \ell$ after the last transaction has been read and processed.

It is easy to see that this invariant is true after the first transaction has been scanned. Suppose now that the invariant is true at the beginning of the $n+1$-th iteration of the while loop, for any n, $0 \leq n \leq |\mathcal{D}| - 1$. We want to show that it will still be true at the end of the $n+1$-th iteration. Let τ be the transaction examined at the $n+1$-th iteration of the loop. If $\tau = |\mathcal{I}|$, the invariant is still true at the end of the $n+1$-th iteration because ℓ does not change and neither does $\mathcal{T}$ because the test of the condition on line 6 of Algorithm 1 fails. The same holds if $|\tau| < \ell$. Consider now the case $|\tau| > \ell$. If $\mathcal{T}$ contained, at the beginning of the $n+1$-th iteration, one transaction equal to τ, then clearly ℓ would not change and neither does $\mathcal{T}$, so the invariant is still true at the end of the $n+1$-th iteration. Suppose now that $|\tau| > \ell$ and that $\mathcal{T}$ did not contain any transaction equal to τ. Let ℓ_i be, for $i = 1, \ldots, |\mathcal{D}| - 1$, the value of ℓ at the end of the i-th iteration, and let $\ell_0 = 1$. If $\mathcal{T}$ contained, at the beginning of the $n+1$-th iteration, zero transactions of length ℓ_n, then necessarily it contained ℓ_n transactions of length greater than ℓ_n, by our assumption that the invariant was true at the end of the n-th iteration. Since $|\tau| > \ell_n$, it follows that $\mathcal{R} = \mathcal{T} \cup \{\tau\}$ contains $\ell_n + 1$ transactions of size at least $\ell_n + 1$; hence, the algorithm at the end of the $n+1$-th iteration has seen $\ell_n + 1$ transactions of length at least $\ell_n + 1$, so $\ell = \ell_{n+1} = \ell_n + 1$. This implies that, at the end of iteration $n+1$, the set $\mathcal{T}$ must have size $\ell_{n+1} = \ell_n + 1$; that is, it must contain

one transaction more than at the beginning of the $n+1$-th iteration. This is indeed the case because the value q computed on line 8 of Algorithm 1 is exactly $|\mathcal{R}| = \ell_n + 1$ because of what we just argued about $\mathcal{R}$. Therefore, $\mathcal{T}$ is exactly $\mathcal{R}$ at the end of the $n+1$-th iteration and contains the $\ell = \ell_{n+1}$ longest different transactions of length at least ℓ, which is exactly what is expressed by the invariant. If instead $\mathcal{T}$ contained, at the beginning of the $n+1$-th iteration, one or more transactions of length ℓ_n, then $\mathcal{T}$ contains at most $\ell_n - 1$ transactions of length greater than ℓ_n, and $\mathcal{R}$ contains at most ℓ_n transactions of length at least $\ell_n + 1$; hence, $q = \ell_n$. This also means that the algorithm has seen, before the beginning of the $n+1$-th iteration, at most $\ell_n - 1$ different transactions strictly longer than ℓ_n. Hence, after seeing τ, the algorithm has seen at most ℓ_n transactions of length at least $\ell_n + 1$, so at the end of the $n+1$-th iteration we will have $\ell = \ell_{n+1} = \ell_n$. This means that the size of $\mathcal{T}$ at the end of the $n+1$-th iteration is the same as it was at the beginning of the same iteration. This is indeed the case because of what we argued about q. At the end of the $n+1$-th iteration, $\mathcal{T}$ contains (1) all transactions of length greater than ℓ_n that it already contained at the end of the n-th iteration, (2) the transaction τ, and (3) all but one transactions of length ℓ_n that it contained at the end of the n-th iteration. Hence, the invariant is true at the end of the $n+1$-th iteration because ℓ did not change, and we replaced in $\mathcal{T}$ a transaction of length ℓ_n with a longer transaction (i.e., τ). Consider now the case of $|\tau| = \ell$. Clearly, if there is a transaction in $\mathcal{T}$ that is equal to τ, the invariant is still true at the end of the $n+1$-th iteration because ℓ does not change, and $\mathcal{T}$ stays the same. If $\mathcal{T}$ did not contain, at the beginning of the $n+1$-th iteration, any transaction equal to τ, then also in this case ℓ would not change (i.e., $\ell = \ell_{n+1} = \ell_n$) because, by definition of ℓ, the algorithm already saw at least ℓ different transactions of length at least ℓ. This implies that $\mathcal{T}$ must have, at the end of the $n+1$-th iteration, the same size that it had at the beginning of the $n+1$-th iteration. This is indeed the case because the set $\mathcal{R}$ contains $\ell + 1$ different transactions of size at least ℓ, but there is no value $b > \ell$ for which $\mathcal{R}$ contains b transactions of length at least b because of what we argued about ℓ; hence, $|\mathcal{T}| = q = \ell$. At the end of the $n+1$-th iteration, the set $\mathcal{T}$ contains (1) all the transactions of length greater than ℓ that it contained at the beginning of the $n+1$-th iteration and (2) enough transactions of length ℓ to make $|\mathcal{T}| = \ell$. This means that $\mathcal{T}$ can contain, at the end of the $n+1$-th iteration, exactly the same set of transactions that it contained at the beginning $n+1$-th iteration. And since, as we argued, ℓ does not change, then the invariant is still true at the end of the $n+1$-th iteration. This completes our proof that the invariant still holds at the end of the $n+1$ iteration for any n and therefore holds at the end of the algorithm, thus proving the thesis. □

The fact that the computation of the d-bound can be easily performed with a single linear scan of the dataset in an online greedy fashion makes it extremely practical also for updating the bound as new transactions are added to the dataset.

4.2. Speeding up the VC-dimension Approximation Task

We showed that the computation of the d-index or of the d-bound can be efficiently performed and that especially the latter only requires a single linear scan of the dataset in a block-by-block fashion if the dataset is stored on disk. In some settings, this may still be an expensive operation. We now present two ways to reduce the cost of this operation.

Empirical VC-dimension. The *empirical VC-dimension* of a range space $S = (X, R)$ on a subset $Y \subseteq X$ of the set of points is the VC-dimension of the range space (Y, R'), where $R' = \{Y \cap f : f \in R\}$ [Boucheron et al. 2005, Sect. 3]. If Y is a random sample from X of size ℓ, and the empirical VC-dimension of S on Y is bounded above by v',

then with probability at least $1 - \delta$, Y is an ε-approximation for X for

$$\varepsilon = 2\sqrt{\frac{2v' \log(\ell + 1)}{\ell}} + \sqrt{\frac{2 \log \frac{2}{\delta}}{\ell}}. \tag{4}$$

Thus, it is possible to create a random sample $\mathcal{S}$ of the dataset $\mathcal{D}$ of the desired size $|\mathcal{S}|$, compute the d-index or the d-bound *on the sample* (which is less expensive than computing it on the whole dataset and, for the d-bound, can be done while creating $\mathcal{S}$), and, finally, after fixing δ, use Equation (4) to compute the ε for which $\mathcal{S}$ is an ε-approximation. Thus, we have a faster method for estimating the VC-dimension that, as we will show in the Section 5, can be used to extract an absolute ε-close approximation to the collection of (top-K) FIs and ARs.

Estimating the d-index from the Transaction Length Distribution. When a Bayesian approach is justified, one views the dataset $\mathcal{D}$ as a sample of n transactions generated by a random process with some known (or assumed) priors. A number of mixture models have been proposed in the literature for modeling dataset generation; the most commonly used is the Dirichlet Process Mixture model [He and Shapiro 2012]. In general, we assume that the generating process $\pi_{\bar{\alpha}}$ belongs to a known parametrized family of distributions $\Pi(\alpha)$ where α represents the parameters of the distribution. Deriving the parameter $\bar{\alpha}$ corresponding to the distribution of transaction lengths according to which the dataset $\mathcal{D}$ was generated can be done by sampling transactions from $\mathcal{D}$ and using techniques for parameter estimation for a distribution from $\Pi(\alpha)$ [Lehmann and Casella 1998; Hastie et al. 2009]. Once the parameter $\bar{\alpha}$ is (probabilistically) known, an upper bound b to the d-index d can be easily derived (probabilistically). Estimating the parameter $\bar{\alpha}$ through sampling may take less time than performing a scan of the entire dataset to compute the d-bound (especially when the dataset is very large): a fast sequential sampling algorithm like Vitter's Method D [Vitter 1987] is used, and the estimation procedure is fast.

4.3. Connection with Monotone Monomials

There is an interesting connection between itemsets built on a ground set $\mathcal{I}$ and the class of *monotone monomials on* $|\mathcal{I}|$ *literals*. A *monotone* monomial is a conjunction of literals with no negations. The class MONOTONE-MONOMIALS$_{|\mathcal{I}|}$ is the class of all monotone monomials on $|\mathcal{I}|$ Boolean variables, including the constant functions **0** and **1**. The VC-dimension of the range space

$$\left(\{0,1\}^{|\mathcal{I}|}, \text{MONOTONE-MONOMIALS}_{|\mathcal{I}|}\right)$$

is exactly $|\mathcal{I}|$ [Natschläger and Schmitt 1996, Coroll. 3]. It is easy to see that it is always possible to build a bijective map between the itemsets in $2^{\mathcal{I}}$ and the elements of MONOTONE-MONOMIALS$_{|\mathcal{I}|}$ and that transactions built on the items in $\mathcal{I}$ correspond to points of $\{0,1\}^{|\mathcal{I}|}$. This implies that a dataset $\mathcal{D}$ can be seen as a sample from $\{0,1\}^{|\mathcal{I}|}$.

Solving the problems we are interested in by using the VC-dimension $|\mathcal{I}|$ of monotone-monomials as an upper bound to the VC-dimension of the itemsets would have resulted in a much larger sample size than sufficient, given that $|\mathcal{I}|$ can be much larger than the d-index of a dataset. Instead, the VC-dimension of the range space $(\mathcal{D}, \boldsymbol{R})$ associated to a dataset $\mathcal{D}$ is equivalent to the VC-dimension of the range space $(\mathcal{D}, \text{MONOTONE-MONOMIALS}_{|\mathcal{I}|})$, which is the *empirical VC-Dimension* of the range space $(\{0,1\}^{|\mathcal{I}|}, \text{MONOTONE-MONOMIALS}_{|\mathcal{I}|})$ measured on $\mathcal{D}$. Our results, therefore, also show a tight bound to the empirical VC-dimension of the class of monotone monomials on $|\mathcal{I}|$ variables.

5. MINING (TOP-K) FREQUENT ITEMSETS AND ASSOCIATION RULES

We apply the VC-dimension results to constructing efficient sampling algorithms with performance guarantees for approximating the collections of FIs, top-K FIs, and ARs.

5.1. Mining Frequent Itemsets

We construct bounds for the sample size needed to obtain relative/absolute ε-close approximations to the collection of FIs. The algorithms to compute the approximations use a standard exact FI mining algorithm on the sample, with an appropriately adjusted minimum frequency threshold, as formalized in the following lemma.

LEMMA 5.1. *Let $\mathcal{D}$ be a dataset with transactions built on a ground set $\mathcal{I}$, and let v be the VC-dimension of the range space associated to $\mathcal{D}$. Let $0 < \varepsilon, \delta < 1$. Let $\mathcal{S}$ be a random sample of $\mathcal{D}$ with size*

$$|\mathcal{S}| = \min\left\{|\mathcal{D}|, \frac{4c}{\varepsilon^2}\left(v + \log\frac{1}{\delta}\right)\right\},$$

for some absolute constant c. Then, $\mathsf{FI}(\mathcal{S}, \mathcal{I}, \theta - \varepsilon/2)$ is an absolute ε-close approximation to $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$ with probability at least $1 - \delta$.

PROOF. Suppose that $\mathcal{S}$ is an $\varepsilon/2$-approximation of the range space (X, R) corresponding to $\mathcal{D}$. From Theorem 3.8, we know that this happens with probability at least $1 - \delta$. This means that for all $X \subseteq \mathcal{I}$, $f_{\mathcal{S}}(X) \in [f_{\mathcal{D}}(X) - \varepsilon/2, f_{\mathcal{D}}(X) + \varepsilon/2]$. This holds in particular for the itemsets in $\mathcal{C} = \mathsf{FI}(\mathcal{S}, \mathcal{I}, \theta - \varepsilon/2)$, which therefore satisfies Property 3 from Definition 3.3. It also means that for all $X \in \mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$, $f_{\mathcal{S}}(X) \geq \theta - \varepsilon/2$, so $\mathcal{C}$ also guarantees Property 1 from Definition 3.3. Now, let $Y \subseteq \mathcal{I}$ be such that $f_{\mathcal{D}}(Y) < \theta - \varepsilon$. Then, for the properties of $\mathcal{S}$, $f_{\mathcal{S}}(Y) < \theta - \varepsilon/2$ (i.e., $Y \notin \mathcal{C}$), which allows us to conclude that $\mathcal{C}$ also has Property 2 from Definition 3.3. □

We stress again that here and in the following theorems, the constant c is absolute and does not depend on $\mathcal{D}$ or on d, ε, or δ.

One very interesting consequence of this result is that we do not need to know in advance the minimum frequency threshold θ in order to build the sample: The properties of the ε-approximation allow us to use the same sample for any threshold and for different thresholds (i.e., the sample does not need to be rebuilt if we want to mine it with a threshold θ first and with another threshold θ' later).

It is important to note that the VC-dimension associated to a dataset, and therefore the sample size from Equation (2) needed to probabilistically obtain an ε-approximation, is independent from the size (number of transactions) in $\mathcal{D}$ and also of the size of $\mathsf{FI}(\mathcal{S}, \mathcal{I}, \theta)$. It is also always smaller or at most as large as the d-index d, which is always less than or equal to the length of the longest transaction in the dataset, which in turn is less than or equal to the number of different items $|\mathcal{I}|$.

To obtain a relative ε-close approximation, we need to add a dependency on θ, as shown in the following lemma.

LEMMA 5.2. *Let $\mathcal{D}$, v, ε, and δ be as in Lemma 5.1. Let $\mathcal{S}$ be a random sample of $\mathcal{D}$ with size*

$$|\mathcal{S}| = \min\left\{|\mathcal{D}|, \frac{4(2+\varepsilon)c}{\varepsilon^2\theta(2-\varepsilon)}\left(v\log\frac{2+\varepsilon}{\theta(2-\varepsilon)} + \log\frac{1}{\delta}\right)\right\},$$

for some absolute absolute constant c. Then, $\mathsf{FI}(\mathcal{S}, \mathcal{I}, (1 - \varepsilon/2)\theta)$ is a relative ε-close approximation to $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$ with probability at least $1 - \delta$.

PROOF. Let $p = \theta(2-\varepsilon)/(2+\varepsilon)$. From Theorem 3.8, the sample $\mathcal{S}$ is a relative $(p, \varepsilon/2)$-approximation of the range space associated to $\mathcal{D}$ with probability at least $1-\delta$. For any

itemset X in $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$, we have $f_{\mathcal{D}}(X) \geq \theta > p$, so $f_{\mathcal{S}}(X) \geq (1-\varepsilon/2) f_{\mathcal{D}}(X) \geq (1-\varepsilon/2)\theta$, which implies $X \in \mathsf{FI}(\mathcal{S}, \mathcal{I}, (1-\varepsilon/2)\theta))$, so Property 1 from Definition 3.3 holds. Now let X be an itemset with $f_{\mathcal{D}}(X) < (1-\varepsilon)\theta$. From our choice of p, we always have $p > (1-\varepsilon)\theta$, so $f_{\mathcal{S}}(X) \leq p(1+\varepsilon/2) < \theta(1-\varepsilon/2)$. This means $X \notin \mathsf{FI}(\mathcal{S}, \mathcal{I}, (1-\varepsilon/2)\theta))$, as requested by Property 2 from Definition 3.3. Since $(1-\varepsilon/2)\theta = p(1+\varepsilon/2)$, it follows that only itemsets X with $f_{\mathcal{D}}(X) \geq p$ can be in $\mathsf{FI}(\mathcal{S}, \mathcal{I}, (1-\varepsilon/2)\theta))$. For these itemsets, it holds $|f_{\mathcal{S}}(X) - f_{\mathcal{D}}(X)| \leq f_{\mathcal{D}}(X)\varepsilon/2$, as requested by Property 3 from Definition 3.3. □

5.2. Mining Top-*K* Frequent Itemsets

Given the equivalence $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K) = \mathsf{FI}(\mathcal{D}, \mathcal{I}, f_{\mathcal{D}}^{(K)})$, we could use the FI sampling algorithms just described if we had a good approximation of $f_{\mathcal{D}}^{(K)}$, the threshold frequency of the top-K FIs.

For the absolute ε-close approximation, we first execute a standard top-K FI mining algorithm on the sample to estimate $f_{\mathcal{D}}^{(K)}$ and then run a standard FI mining algorithm on the same sample using a minimum frequency threshold depending on our estimate of $f_{\mathcal{S}}^{(K)}$. Lemma 5.3 formalizes this intuition.

LEMMA 5.3. *Let $\mathcal{D}$, v, ε, and δ be as in Lemma 5.1. Let K be a positive integer. Let $\mathcal{S}$ be a random sample of $\mathcal{D}$ with size*

$$|\mathcal{S}| = \min\left\{|\mathcal{D}|, \frac{16c}{\varepsilon^2}\left(v + \log\frac{1}{\delta}\right)\right\},$$

for some absolute constant c, then $\mathsf{FI}(\mathcal{S}, \mathcal{I}, f_{\mathcal{S}}^{(K)} - \varepsilon/2)$ is an absolute ε-close approximation to $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K)$ with probability at least $1-\delta$.

PROOF. Suppose that $\mathcal{S}$ is an $\varepsilon/4$-approximation of the range space (X, R) corresponding to $\mathcal{D}$. From Theorem 3.8, we know that this happens with probability at least $1-\delta$. Thus, for all $Y \subseteq \mathcal{I}$, $f_{\mathcal{S}}(Y) \in [f_{\mathcal{D}}(Y) - \varepsilon/4, f_{\mathcal{D}}(Y) + \varepsilon/4]$. Consider now $f_{\mathcal{S}}^{(K)}$, the frequency of the K-th most FI in the sample. Clearly, $f_{\mathcal{S}}^{(K)} \geq f_{\mathcal{D}}^{(K)} - \varepsilon/4$, because there are at least K itemsets (e.g., any subset of size K of $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K)$) with frequency in the sample at least $f_{\mathcal{D}}^{(K)} - \varepsilon/4$. On the other hand, $f_{\mathcal{S}}^{(K)} \leq f_{\mathcal{D}}^{(K)} + \varepsilon/4$ because there cannot be K itemsets with a frequency in the sample greater than $f_{\mathcal{D}}^{(K)} + \varepsilon/4$: Only itemsets with frequency in the dataset strictly greater than $f_{\mathcal{D}}^{(K)}$ can have a frequency in the sample greater than $f_{\mathcal{D}}^{(K)} + \varepsilon/4$, and there are at most $K-1$ such itemsets. Now let $\eta = f_{\mathcal{S}}^{(K)} - \varepsilon/2$, and consider $\mathsf{FI}(\mathcal{S}, \mathcal{I}, \eta)$. We have $\eta \leq f_{\mathcal{D}}^{(K)} - \varepsilon/4$, so for the properties of $\mathcal{S}$, $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K) = \mathsf{FI}(\mathcal{D}, \mathcal{I}, f_{\mathcal{D}}^{(K)}) \subseteq \mathsf{FI}(\mathcal{S}, \mathcal{I}, \eta)$, which then guarantees Property 1 from Definition 3.3. On the other hand, let Y be an itemset such that $f_{\mathcal{D}}(Y) < f_{\mathcal{D}}^{(K)} - \varepsilon$. Then, $f_{\mathcal{S}}(Y) < f_{\mathcal{D}}^{(K)} - 3\varepsilon/4 \leq \eta$, so $Y \notin \mathsf{FI}(\mathcal{S}, \mathcal{I}, \eta)$, corresponding to Property 2 from Definition 3.3. Property 3 from Definition 3.3 follows from the properties of $\mathcal{S}$. □

Note that, as in the case of the sample size required for an absolute ε-close approximation to $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$, we do not need to know K in advance to compute the sample size for obtaining an absolute ε-close approximation to $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K)$.

Two different samples are needed for computing a relative ε-close approximation to $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K)$, the first one to compute a lower bound to $f_{\mathcal{D}}^{(K)}$, the second to extract the approximation. Details for this case are presented in Lemma 5.4.

LEMMA 5.4. *Let $\mathcal{D}$, v, ε, and δ be as in Lemma 5.1. Let K be a positive integer. Let δ_1, δ_2 be two reals such that $(1-\delta_1)(1-\delta_2) \geq (1-\delta)$. Let $\mathcal{S}_1$ be a random sample of $\mathcal{D}$*

with some size

$$|\mathcal{S}_1| = \frac{\phi c}{\varepsilon^2}\left(v + \log\frac{1}{\delta_1}\right)$$

for some $\phi > 2\sqrt{2}/\varepsilon$ *and some absolute constant* c. *If* $f^{(K)}_{\mathcal{S}_1} \geq (2\sqrt{2})/(\varepsilon\phi)$, *then let* $p = (2-\varepsilon)\theta/(2+\varepsilon)$ *and let* $\mathcal{S}_2$ *be a random sample of* $\mathcal{D}$ *of size*

$$|\mathcal{S}_2| = \min\left\{|\mathcal{D}|, \frac{4c}{\varepsilon^2 p}\left(v\log\frac{1}{p} + \log\frac{1}{\delta}\right)\right\}$$

for some absolute constant c. *Then,* $\mathsf{FI}(\mathcal{S}_2, \mathcal{I}, (1-\varepsilon/2)(f^{(K)}_{\mathcal{S}_1} - \varepsilon/\sqrt{2\phi}))$ *is a relative* ε*-close approximation to* $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K)$ *with probability at least* $1-\delta$.

Proof. Assume that $\mathcal{S}_1$ is a $\varepsilon/\sqrt{2\phi}$-approximation for $\mathcal{D}$ and $\mathcal{S}_2$ is a relative $(p, \varepsilon/2)$-approximation for $\mathcal{D}$. The probability of these two events happening at the same time is at least $1-\delta$, from Theorem 3.8.

Following the steps of the proof of Lemma 5.3, we can easily get that, from the properties of $\mathcal{S}_1$,

$$f^{(K)}_{\mathcal{S}_1} - \frac{\varepsilon}{\sqrt{2\phi}} \leq f^{(K)}_{\mathcal{D}} \leq f^{(K)}_{\mathcal{S}_1} + \frac{\varepsilon}{\sqrt{2\phi}}. \tag{5}$$

Consider now an element $X \in \mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K)$. We have by definition $f_{\mathcal{D}}(X) \geq f^{(K)}_{\mathcal{D}} > f^{(K)}_{\mathcal{S}_1} - \varepsilon/\sqrt{2\phi} \geq p$, and, from the properties of $\mathcal{S}_2$, it follows that $f_{\mathcal{S}}(X) \geq (1-\varepsilon/2) f_{\mathcal{D}}(X) \geq (1-\varepsilon/2)(f^{(K)}_{\mathcal{S}_1} - \varepsilon/\sqrt{2\phi})$, which implies $X \in \mathsf{FI}(\mathcal{S}_2, \mathcal{I}, (1-\varepsilon/2)(f^{(K)}_{\mathcal{S}_1} - \varepsilon/\sqrt{2\phi}))$ and therefore Property 1 from Definition 3.3 holds for $FI(\mathcal{S}_2, \mathcal{I}, \eta)$.

Now let Y be an itemset such that $f_{\mathcal{D}}(Y) < (1-\varepsilon) f^{(K)}_{\mathcal{D}}$. From our choice of p, we have that $f_{\mathcal{D}}(Y) < p$. Then, $f_{\mathcal{S}_2}(Y) < (1+\varepsilon/2)p < (1-\varepsilon/2)(f^{(K)}_{\mathcal{S}_1} - \varepsilon/\sqrt{2\phi})$. Therefore, $Y \notin \mathsf{FI}(\mathcal{S}_2, \mathcal{I}, \eta)$, and Property 2 from Definition 3.3 is guaranteed.

Property 3 from Definition 3.3 follows from Equation (5) and the properties of $\mathcal{S}_2$. □

5.3. Mining Association Rules

Our final theoretical contribution concerns the discovery of relative/absolute approximations to $\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \eta)$ from a sample. Lemma 5.5 builds on a result from Chakaravarthy et al. [2009, Sect. 5] and covers the *relative* case, whereas Lemma 5.6 deals with the *absolute* one.

Lemma 5.5. *Let* $0 < \delta, \varepsilon, \theta, \gamma < 1$, $\phi = \max\{2+\varepsilon, 2-\varepsilon+2\sqrt{1-\varepsilon}\}$, $\eta = \varepsilon/\phi$, *and* $p = \theta(1-\eta)/(1+\eta)$. *Let* $\mathcal{D}$ *be a dataset and* v *be the VC-dimension of the range space associated to* $\mathcal{D}$. *Let* $\mathcal{S}$ *be a random sample of* $\mathcal{D}$ *of size*

$$|\mathcal{S}| = \min\left\{|\mathcal{D}|, \frac{c}{\eta^2 p}\left(v\log\frac{1}{p} + \log\frac{1}{\delta}\right)\right\} \tag{6}$$

for some absolute constant c. *Then,* $\mathsf{AR}(\mathcal{S}, \mathcal{I}, (1-\eta)\theta, \gamma(1-\eta)/(1+\eta))$ *is a relative* ε*-close approximation to* $\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ *with probability at least* $1-\delta$.

Proof. Suppose $\mathcal{S}$ is a relative (p, η)-approximation for the range space corresponding to $\mathcal{D}$. From Theorem 3.8 we know this happens with probability at least $1-\delta$.

Let $W \in \mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ be the association rule "$A \Rightarrow B$", where A and B are itemsets. By definition $f_{\mathcal{D}}(W) = f_{\mathcal{D}}(A \cup B) \geq \theta > p$. From this and the properties of $\mathcal{S}$, we get

$$f_{\mathcal{S}}(W) = f_{\mathcal{S}}(A \cup B) \geq (1-\eta) f_{\mathcal{D}}(A \cup B) \geq (1-\eta)\theta.$$

Note that, from the fact that $f_{\mathcal{D}}(W) = f_{\mathcal{D}}(A \cup B) \geq \theta$, it follows that $f_{\mathcal{D}}(A)$, $f_{\mathcal{D}}(B) \geq \theta > p$, for the antimonotonicity property of the frequency of itemsets.

By definition, $c_{\mathcal{D}}(W) = f_{\mathcal{D}}(W)/f_{\mathcal{D}}(A) \geq \gamma$. Then,

$$c_{\mathcal{S}}(W) = \frac{f_{\mathcal{S}}(W)}{f_{\mathcal{S}}(A)} \geq \frac{(1-\eta)f_{\mathcal{D}}(W)}{(1+\eta)f_{\mathcal{D}}(A)} \geq \frac{1-\eta}{1+\eta} \cdot \frac{f_{\mathcal{D}}(W)}{f_{\mathcal{D}}(A)} \geq \frac{1-\eta}{1+\eta}\gamma.$$

It follows that $W \in \mathsf{AR}(\mathcal{S}, \mathcal{I}, (1-\eta)\theta, \gamma(1-\eta)/(1+\eta))$; hence, Property 1 from Definition 3.4 is satisfied.

Now let Z be the association rule "$C \Rightarrow D$", such that $f_{\mathcal{D}}(Z) = f_{\mathcal{D}}(C \cup D) < (1-\varepsilon)\theta$. But from our definitions of η and p, it follows that $f_{\mathcal{D}}(Z) < p < \theta$; hence, $f_{\mathcal{S}}(Z) < (1+\eta)p < (1-\eta)\theta$, and therefore $Z \notin \mathsf{AR}(\mathcal{S}, \mathcal{I}, (1-\eta)\theta, \gamma(1-\eta)(1+\eta))$, as requested by Property 2 from Definition 3.4.

Consider now an association rule $Y =$ "$E \Rightarrow F$" such that $c_{\mathcal{D}}(Y) < (1-\varepsilon)\gamma$. Clearly, we are only concerned with Y such that $f_{\mathcal{D}}(Y) \geq p$; otherwise, we just showed that Y cannot be in $\mathsf{AR}(\mathcal{S}, \mathcal{I}, (1-\eta)\theta, \gamma(1-\eta)/(1+\eta))$. From this and the antimonotonicity property, it follows that $f_{\mathcal{D}}(E)$, $f_{\mathcal{D}}(F) \geq p$. Then,

$$c_{\mathcal{S}}(Y) = \frac{f_{\mathcal{S}}(Y)}{f_{\mathcal{S}}(E)} \leq \frac{(1+\eta)f_{\mathcal{D}}(Y)}{(1-\eta)f_{\mathcal{D}}(E)} < \frac{1+\eta}{1-\eta}(1-\varepsilon)\gamma < \frac{1-\eta}{1+\eta}\gamma,$$

where the last inequality follows from the fact that $(1-\eta)^2 > (1+\eta)(1-\varepsilon)$ for our choice of η. We can conclude that $Y \notin \mathsf{AR}(\mathcal{S}, \mathcal{I}, (1-\varepsilon)\theta, \gamma(1-\eta)/(1+\eta)\gamma)$ and therefore Property 4 from Definition 3.4 holds.

Properties 3 and 5 from Definition 3.4 follow from these steps (i.e., what association rules can be in the approximations), from the definition of ϕ, and from the properties of $\mathcal{S}$. □

LEMMA 5.6. *Let $\mathcal{D}$, v, θ, γ, ε, and δ be as in Lemma 5.5 and let $\varepsilon_{\text{rel}} = \varepsilon/\max\{\theta, \gamma\}$. Fix $\phi = \max\{2+\varepsilon, 2-\varepsilon_{\text{rel}} + 2\sqrt{1-\varepsilon_{\text{rel}}}\}$, $\eta = \varepsilon_{\text{rel}}/\phi$, and $p = \theta(1-\eta)/(1+\eta)$. Let $\mathcal{S}$ be a random sample of $\mathcal{D}$ of size*

$$|\mathcal{S}| = \min\left\{|\mathcal{D}|, \frac{c}{\eta^2 p}\left(v \log\frac{1}{p} + \log\frac{1}{\delta}\right)\right\} \tag{7}$$

for some absolute constant c. Then $\mathsf{AR}(\mathcal{S}, \mathcal{I}, (1-\eta)\theta, \gamma(1-\eta)/(1+\eta))$ is an absolute ε-close approximation to $\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$.

PROOF. The thesis follows from Lemma 5.5 by setting ε there to ε_{rel}. □

Note that the sample size needed for absolute ε-close approximations to $\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ depends on θ and γ, which was not the case for absolute ε-close approximations to $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$ and $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K)$.

5.4. Other Interestingness Measures

Confidence is not the only measure for the interestingness of an association rule. Other measures include lift, IS (cosine), all-confidence, Jaccard index, leverage, conviction, and many more [Tan et al. 2004]. In this section, we apply our general technique to obtain good approximations with respect to a number of these measures while also showing the limitation of our technique with respect to other criteria.

We use the term "absolute" (or "relative") ε-close approximation as defined in Definition 3.4, appropriately adapted to the relevant measure in place of the confidence. We also extend our notation and denote the collection of ARs with frequency at least θ and interestingness at least γ according to a measure w by $\mathsf{AR}_w(\mathcal{D}, \mathcal{I}, \theta, \gamma)$; that is, indicating the measure in the subscript of "AR."

The first two measures we deal with are *all-confidence* and *IS* (also known as *Cosine*). They are defined as follows:

$$\text{all-confidence: } ac_{\mathcal{D}}(A \Rightarrow B) = \frac{f_{\mathcal{D}}(A \cup B)}{max_{a \in A \cup B} f_{\mathcal{D}}(A)}$$

$$\text{IS (Cosine): } is_{\mathcal{D}}(A \Rightarrow B) = \frac{f_{\mathcal{D}}(A \cup B)}{\sqrt{f_{\mathcal{D}}(A) f_{\mathcal{D}}(B)}}$$

Because the approximation errors in the numerators and denominators of these measures are the same as in computing the confidence, we can follow exactly the same steps as in the proof of Lemmas 5.5 and 5.6 and obtain the same procedures, parameters, and sample sizes from Equations (6) and (7) to extract relative and absolute ε-close approximations to the collection of ARs according to these measures.

Lift. The *lift* of an association rule "$A \Rightarrow B$" is defined as

$$\ell_{\mathcal{D}}(A \Rightarrow B) = \frac{f_{\mathcal{D}}(A \cup B)}{f_{\mathcal{D}}(A) f_{\mathcal{D}}(B)}.$$

We have the following result about computing a relative ε-close approximation to the collection of ARs according to lift.

LEMMA 5.7. *Let $\mathcal{D}$, v, θ, γ, ε, and δ be as in Lemma 5.5. There exists a value η such that, if we let $p = \theta(1-\eta)/(1+\eta)$, and let $\mathcal{S}$ be random sample of $\mathcal{D}$ of size*

$$|\mathcal{S}| = \min\left\{|\mathcal{D}|, \frac{c}{\eta^2 p}\left(v \log\frac{1}{p} + \log\frac{1}{\delta}\right)\right\}$$

for some absolute constant c, we have that $\mathsf{AR}_\ell(\mathcal{S}, \mathcal{I}, (1-\eta)\theta, \gamma(1-\eta)/(1+\eta))$ is a relative ε-close approximation to $\mathsf{AR}_\ell(\mathcal{D}, \mathcal{I}, \theta, \gamma)$.

PROOF. In order for $\mathsf{AR}_\ell(\mathcal{S}, \mathcal{I}, (1-\eta)\theta, \gamma(1-\eta)/(1+\eta))$ to satisfy the properties of a relative ε-close approximation, η must be a solution to the following system of inequalities:

$$\begin{cases} (1-\varepsilon)(1+\eta)^3 < (1-\eta)^3 \\ \dfrac{1+\eta}{(1-\eta)^2} \leq 1+\varepsilon \\ \dfrac{1-\eta}{(1+\eta)^2} \geq 1-\varepsilon \\ 0 \leq \eta < 1 \end{cases}$$

The first inequality expresses the requirement of Property 4 from Definition 3.4. The second and third inequalities deal with Properties 1, 3, and 5. The last inequality limits the domain of η. Property 2 from Definition 3.4 would be enforced by the choice of p. It can be verified that this system admits solutions. Once the value of η has been determined, we can proceed as in the proof of Lemma 5.5 to prove that all properties from Definition 3.4 are satisfied. □

We can get a result about *absolute* ε-close approximation to $\mathsf{AR}_\ell(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ by following the same derivation of Lemma 5.6.

Piatetsky-Shapiro Measure (Leverage). Another measure of interestingness is the *Piatetsky-Shapiro* measure (also known as *leverage*):

$$ps_{\mathcal{D}}(A \Rightarrow B) = f_{\mathcal{D}}(A \cup B) - f(A) f(B).$$

We first prove that it is possible to obtain an *absolute* ε-close approximation to the collection of ARs according to this measure and then argue that our methods cannot be used to obtain a *relative* ε-close approximation to such collection.

LEMMA 5.8. *Let $\mathcal{D}$, v, θ, γ, ε, and δ be as in Lemma 5.5. Let $\mathcal{S}$ be a random sample of $\mathcal{D}$ of size*

$$|S| = \min\left\{|D|, \frac{64c}{\varepsilon^2}\left(v + \log\frac{1}{\delta}\right)\right\}$$

for some absolute constant c. Then $\mathsf{AR}_{ps}(\mathcal{S}, \mathcal{I}, \theta - \varepsilon/8, \gamma - \varepsilon/2)$ is an absolute ε-close approximation to $\mathsf{AR}_{ps}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ with probability at least $1 - \delta$.

PROOF. Assume that $\mathcal{S}$ is an $\varepsilon/8$-approximation for $\mathcal{D}$. From Theorem 3.8, we know this happens with probability at least $1 - \delta$. This implies that for any itemeset $A \subseteq \mathcal{I}$, we have $|f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A)| \leq \varepsilon/8$, which holds in particular for the association rules in $\mathsf{AR}_{ps}(\mathcal{S}, \mathcal{I}, \theta - \varepsilon/8, \gamma - \varepsilon/2)$; so, Property 3 from Definition 3.4 is satisfied.

Consider now an association rule $W =$ "$A \Rightarrow B$". We have

$$\begin{aligned} ps_{\mathcal{S}}(W) &= f_{\mathcal{S}}(A \cup B) - f_{\mathcal{S}}(A) f_{\mathcal{S}}(B) \geq f_{\mathcal{D}}(A \cup B) - \frac{\varepsilon}{8} - \left(f_{\mathcal{D}}(A) + \frac{\varepsilon}{8}\right)\left(f_{\mathcal{D}}(B) + \frac{\varepsilon}{8}\right) \\ &\geq f_{\mathcal{D}}(A \cup B) - f_{\mathcal{D}}(A) f_{\mathcal{D}}(B) - \frac{\varepsilon}{8}\left(1 + f_{\mathcal{D}}(A) + f_{\mathcal{D}}(B) + \frac{\varepsilon}{8}\right) \\ &\leq ps_{\mathcal{D}}(W) - \frac{\varepsilon}{2}. \end{aligned} \tag{8}$$

We also have:

$$\begin{aligned} ps_{\mathcal{S}}(W) &= f_{\mathcal{S}}(A \cup B) - f_{\mathcal{S}}(A) f_{\mathcal{S}}(B) \leq f_{\mathcal{D}}(A \cup B) + \frac{\varepsilon}{8} - \left(f_{\mathcal{D}}(A) - \frac{\varepsilon}{8}\right)\left(f_{\mathcal{D}}(B) - \frac{\varepsilon}{8}\right) \\ &\leq f_{\mathcal{D}}(A \cup B) - f_{\mathcal{D}}(A) f_{\mathcal{D}}(B) + \frac{\varepsilon}{8}\left(1 + f_{\mathcal{D}}(A) + f_{\mathcal{D}}(B) - \frac{\varepsilon}{8}\right) \\ &\leq ps_{\mathcal{D}}(W) + \frac{\varepsilon}{2} \end{aligned} \tag{9}$$

From Equations (8) and (9), we get that for any association rule W, we have $|ps_{\mathcal{D}}(W) - ps_{\mathcal{S}}(W)| < \varepsilon$; hence, Property 5 from Definition 3.4 holds.

If $W \in \mathsf{AR}_{ps}(\mathcal{S}, \mathcal{I}, \theta, \gamma)$, Equation (8) implies that $W \in \mathsf{AR}_{ps}(\mathcal{S}, \mathcal{I}, \theta - \varepsilon/2, \gamma - \varepsilon/2)$; therefore, Property 1 from Definition 3.4 is satisfied.

Now let Z be an association rule with frequency $f_{\mathcal{D}}(Z) < \theta - \varepsilon$. From the property of $\mathcal{S}$, we have that $f_{\mathcal{S}}(Z) \leq f_{\mathcal{D}}(Z) + \varepsilon/8 < \theta - \varepsilon + \varepsilon/8 < \theta - \varepsilon/8$, so $Z \notin \mathsf{AR}_{ps}(\mathcal{S}, \mathcal{I}, \theta - \varepsilon/8, \gamma - \varepsilon/2)$, which proves Property 2 from Definition 3.4.

Consider now an association rule $Y =$ "$C \Rightarrow D$" with frequency $f_{\mathcal{D}}(Y) > \theta$ but leverage $ps_{\mathcal{D}}(Y) < \gamma - \varepsilon$ ($Y \notin \mathsf{AR}_{ps}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$). From Equation (9), we get that $ps_{\mathcal{S}}(Y) < \gamma - \varepsilon/2$, which implies that $Y \notin \mathsf{AR}_{ps}(\mathcal{S}, \mathcal{I}, \theta - \varepsilon/8, \gamma - \varepsilon/2)$, hence proving Property 4 from Definition 3.4. This concludes our proof. □

We now argue that it is not possible, in general, to extend our methods to obtain a relative ε-close approximation to $\mathsf{AR}_{ps}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$. Suppose that there is a parameter λ for which, for any itemset A, we can find a value $\tilde{f}(A)$ such that $(1 - \lambda) f_{\mathcal{D}}(A) \leq \tilde{f}(A) \leq (1 + \lambda) f_{\mathcal{D}}(A)$. Let $\widetilde{ps}(A \Rightarrow B) = \tilde{f}(A \cup B) - \tilde{f}(A)\tilde{f}(B)$. We would like to show that the values $\widetilde{ps}$ cannot be used to obtain a relative ε-close approximation to $\mathsf{AR}_{ps}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ in general. $0 < \varepsilon, \theta, \gamma < 1$. Among the requirement for a relative ε-close approximation, we have that for an AR "$A \rightarrow B$" in the approximation, it must hold $\widetilde{ps}(A \Rightarrow B) \geq$

$(1-\varepsilon)ps_{\mathcal{D}}(A \Rightarrow B)$. We now show that this is not true in general. We have the following:

$$\begin{aligned}\widetilde{ps}(A \Rightarrow B) &\geq (1-\lambda) f_{\mathcal{D}}(A \cup B) - (1+\lambda)^2 f_{\mathcal{D}}(A) f_{\mathcal{D}}(B) \\ &\geq (1-\varepsilon) f_{\mathcal{D}}(A \cup B) - (1-\varepsilon) f_{\mathcal{D}}(A) f_{\mathcal{D}}(B) \\ &\iff (\varepsilon-\lambda) f_{\mathcal{D}}(A \cup B) - (\varepsilon + 2\lambda + \lambda^2) f_{\mathcal{D}}(A) f_{\mathcal{D}}(B) \geq 0.\end{aligned}$$

Clearly, the inequality on the last line may not be true in general. This means that we cannot, in general, obtain a relative ε-close approximation to $\mathsf{AR}_{ps}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ by approximating the frequencies of all itemsets, no matter how good these would be.

Other Measures. For other measures, it may not be possible or straightforward to analytically derive procedures and sample sizes sufficient to extract good approximations of the collection of ARs according to these measures. Nevertheless, most of them express the interestingness of an AR as a function of the frequencies of the itemsets involved in the rule. Because of this, in practice, high-quality approximation of the frequencies of all itemsets should be sufficient to obtain good approximation of the interestedness of a rule and, therefore, good approximation of the collection of ARs.

5.5. Closed Frequent Itemsets

A Closed Frequent Itemset (CFI) is a FI A whose subsets have all the same frequency $f\_Ds(A)$ of A. The collection of CFIs is a lossless compressed representation of the FIs [Calders et al. 2006]. The collection of CFIs is quite sensitive to sampling, as shown by the following example. Consider the dataset

$$\mathcal{D} = \{\{a, b, c\}, \{a\}, \{b\}, \{c\}\}.$$

Suppose that $\theta = 0.5$. Then, $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta) = \{\{a\}, \{b\}, \{c\}\}$, and this is also the collection of CFIs. Consider the sample $\mathcal{S} = \{\{a, b, c\}, \{b\}\}$ of $\mathcal{D}$. We have that

$$\mathsf{FI}(\mathcal{S}, \mathcal{I}, \theta') = \{\{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$$

for any $\theta' \leq \theta$. But the collection of CFIs is $\{\{b\}, \{a, b, c\}\}$, and it is not a superset of the original collection. Thus, in a sample, a superset of an original CFI may become closed instead of the original one. Therefore, given an absolute ε-close approximation F to $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$ (analogously for a relative approximation), one could obtain a superset of the original collection of CFIs by considering, for each CFI $B \in F$, the set of subsets of B whose frequency in F is less than 2ε far from that of B. As was the case for FIs, a single scan of the dataset is then sufficient to filter out spurious candidates that are not CFIs from the so-obtained collection.

5.6. Discussion

In the previous sections, we presented the bounds to the sample sizes as a function of the VC-dimension v of the range space associated to the dataset. As we argued in Section 4, computing the VC-dimension exactly is not a viable option. We therefore introduced the d-index d and the d-bound q as upper bounds to the VC-dimension; these are efficient to compute, as described in Section 4.1. In practice, one would use d or q, rather than v, to obtain the needed sample sizes.

Chakaravarthy et al. [2009] presented bounds to the sample sizes that depend on the length Δ of the longest transaction. It should be clear that $v \leq d \leq q \leq \Delta$, with the first inequality being strict in the worst case (Theorem 4.6). In real datasets, we have that $v \leq d \leq q \ll \Delta$: a single very long transaction has minimal impact on the VC-dimension or its upper bounds. One can envision cases where an anomalous transaction contains most items from $\mathcal{I}$ while all other transactions have constant length. This would drive up the sample size from Chakaravarthy et al. [2009], while the bounds presented in this work would not be impacted by this anomality.

Moreover, in practice, one could expect v to be much smaller than the d-index d. This is due to the fact that the d-index is really a worst-case bound that should only occur in artificial datasets, as should be evident from the proof of Theorem 4.6. It would be very interesting to investigate better methods to estimate the actual VC-dimension of the range space associated to a dataset, rather than upper-bound it with d or q, because this could lead to much smaller sample sizes. The problem of estimating the VC-dimension of learning machines is a fundamental problem in learning, given that analytical computation of the exact value is usually impossible, as it is in our case. Vapnik et al. [1994] and Shao et al. [2000] presented and refined an experimental procedure to estimate the VC-dimension of a learning machines, and McDonald et al. [2011] gave concentration results for such an estimate. This procedure, although applicable to our case under mild conditions, is not very practical. It is very highly time consuming because it requires the creation and analysis of multiple artificial datasets starting from the original one. Developing efficient ways to estimate the VC-dimension of a range space is an interesting research problem, but outside the scope of this work.

We conclude this discussion by noting that all the bounds we presented have a dependency on $1/\varepsilon^2$. This is due to the use of tail bounds dependent on this quantity in the proof of the bound in Equation (2) to the sample size needed to obtain an ε-approximation. Given that the bound in Equation (2) is in general tight up to a constant [Li et al. 2001], there seems to be little room for improvement of the bounds we presented as a function of ε.

6. EXPERIMENTAL EVALUATION

In this section, we present an extensive experimental evaluation of our methods to extract approximations of $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$, $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K)$, and $\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$.

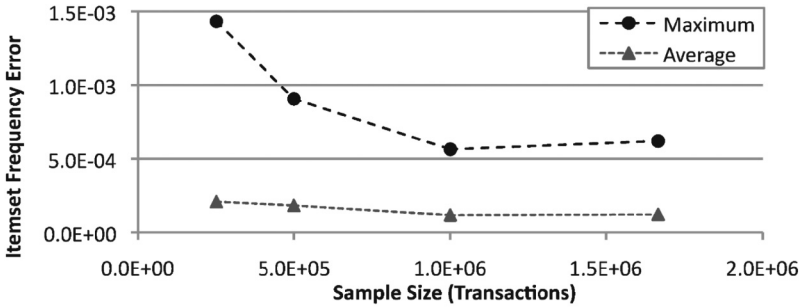
Our first goal is to evaluate the *quality* of the approximations obtained using our methods by comparing the experimental results to the analytical bounds. We also evaluate how strict the bounds are by testing whether the same quality of results can be achieved at sample sizes smaller than those computed by the theoretical analysis. We then show that our methods can significantly speed up the mining process, fulfilling the motivating promises of the use of sampling in the market basket analysis tasks. Last, we compare the sample sizes from our results to the best previous work [Chakaravarthy et al. 2009].

We tested our methods on both real and artificial datasets. The real datasets come from the FIMI'04 repository (http://fimi.ua.ac.be/data/). Because most of them have a moderately small size, we replicated their transactions a number of times, with the only effect being an increase in the size of the dataset but no change in the distribution of the frequencies of the itemsets. The artificial datasets were built such that their corresponding range spaces have a VC-dimension equal to the maximum transaction length, which is the maximum possible, as shown in Theorem 4.5. To create these datasets, we followed the proof of Theorem 4.6 and used the generator included in ARtool (http://www.cs.umb.edu/~laur/ARtool/), which is similar to the one presented in Agrawal and Srikant [1994]. The artificial datasets had 10 million transactions. We used the FP-Growth and Apriori implementations in ARtool to extract FIs and ARs. To compute the d-bound q, which is an upper bound to the d-index d, we used Algorithm 1. In all our experiments, we fixed $\delta = 0.1$. In the experiments involving absolute (resp. relative) ε-close approximations, we set $\varepsilon = 0.01$ (resp. $\varepsilon = 0.05$). The absolute constant c was fixed to 0.5 as estimated by Löffler and Phillips [2009]. This is reasonable because, again, c does not depend in any way on $\mathcal{D}$, ε, δ, the VC-dimension v of the range space, the d-index d or the d-bound q, or any characteristic of the collection of FIs or ARs. No upper bound is currently known for c' when computing the sizes for relative ε-approximations. We used the same value 0.5 for c' and found

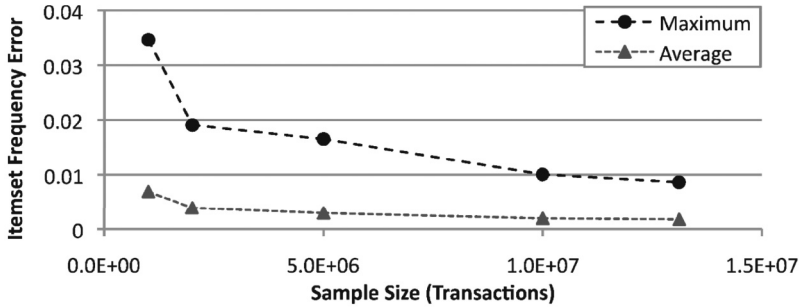
that it worked well in practice. For each dataset, we selected a range of minimum frequency thresholds and a set of values for K when extracting the top-K FIs. For AR discovery, we set the minimum confidence threshold $\gamma \in \{0.5, 0.75, 0.9\}$. For each dataset and each combination of parameters, we created random samples with size as computed by our theorems and with smaller sizes to evaluate the strictness of the bounds. We measured, for each set of parameters, the *absolute frequency error* and the *absolute confidence error*, defined as the error $|f_{\mathcal{D}}(X) - f_{\mathcal{S}}(X)|$ (resp. $|c_{\mathcal{D}}(Y) - c_{\mathcal{S}}(Y)|$) for an itemset X (resp. an association rule Y) in the approximate collection extracted from sample $\mathcal{S}$. When dealing with the problem of extracting *relative* ε-close approximations, we defined the *relative frequency error* to be the absolute frequency error divided by the real frequency of the itemset and, analogously, for the relative confidence error (dividing by the real confidence). In the figures, we plot the maximum and the average for these quantities, taken over all itemsets or ARs in the output collection. In order to limit the influence of a single sample, we computed and plot in the figures the maximum (resp. the average) of these quantities in three runs of our methods on three different samples for each size.

The first important result of our experiments is that, for all problems (FIs, top-K FIs, ARs), for every combination of parameters, and for every run, the collection of itemsets or of ARs obtained using our methods always satisfied the requirements to be an absolute or relative ε-close approximation to the real collection. Thus, in practice, our methods indeed achieve or exceed the theoretical guarantees for approximations of the collections $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$, $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, \theta)$, and $\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$. Given that the collections returned by our algorithms were always a superset of the collections of interest or, in other words, that the *recall* of the collections we returned was always 1.0, we measured the *precision* of the returned collection. In all but one case, this statistic was at least 0.9 (out of a maximum of 1.0), suggesting relatively few false positives in the collection's output. In the remaining case (extracting FIs from the dataset BMS-POS), the precision ranged between 0.59 and 0.8 (respectively, for $\theta = 0.02$ and $\theta = 0.04$). The probability of including an FI or an AR which has a frequency (or confidence, for ARs) of less than θ (or γ) but does not violate the properties of an ε-close approximation and is therefore an "acceptable" false positive depends on the distribution of the real frequencies of the itemsets or ARs in the dataset around the frequency threshold θ (more precisely, below it, within ε or $\varepsilon\theta$): If many patterns have a real frequency in this interval, then it is highly probable that some of them will be included in the collections given in output, driving precision down. Clearly, this probability depends on the number of patterns that have a real frequency close to θ. Given that usually the lower the frequency, the higher the number of patterns with that frequency, this implies that our methods may include more "acceptable" false positives in the output at very low frequency thresholds. Once again, this depends on the distribution of the frequencies and does not violate the guarantees offered by our methods. It is possible to use the output of our algorithms as a set of *candidate patterns* that can be reduced to the real exact output (i.e., with no false positives) with a single scan of the dataset.

Evaluating the strictness of the bounds to the sample size was the second goal of our experiments. In Figure 2(a), we show the behavior of the maximum frequency error as a function of the sample size in the itemsets obtained from samples using the method presented in Lemma 5.1 (i.e., we are looking for an *absolute* ε-close approximation to $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$). The rightmost plotted point corresponds to the sample size computed by the theoretical analysis. We are showing the results for the dataset BMS-POS replicated 40 times (d-index $d = 81$), mined with $\theta = 0.02$. It is clear from the picture that the guaranteed error bounds are achieved even at sample sizes smaller than that computed by the analysis and that the error at the sample size derived from the theory (rightmost plotted point for each line) is one to two orders of magnitude smaller than



(a) Absolute Itemset Frequency Error, BMS-POS dataset, $d = 81$, $\theta = 0.02$, $\varepsilon = 0.01$, $\delta = 0.1$



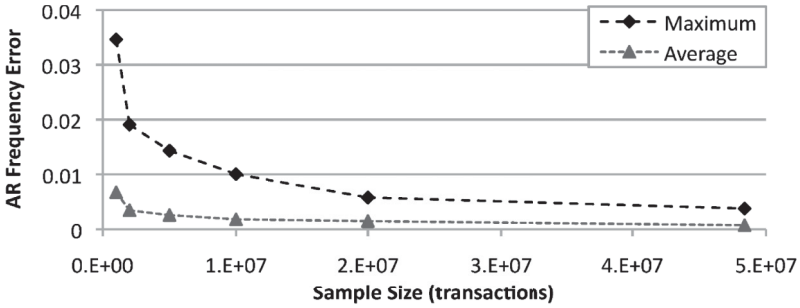
(b) Relative Itemset Frequency Error, artificial dataset, $v = 33$, $\theta = 0.01$, $\varepsilon = 0.05$, $\delta = 0.1$

Fig. 2. Absolute/Relative ε-close approximation to $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$.

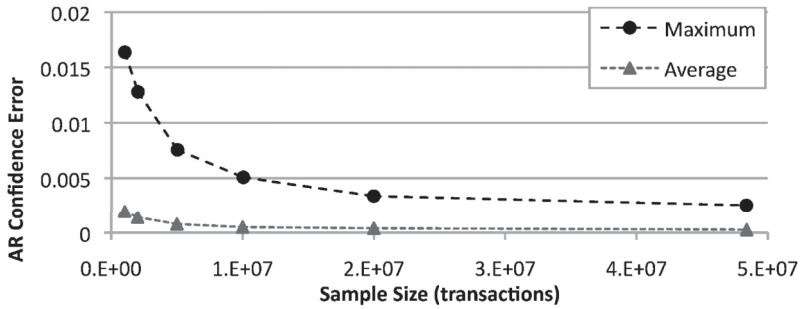
the maximum tolerable error $\varepsilon = 0.01$. This can be explained by the fact that the d-bound used to compute the sample size is, in practice (as we argued in Section 5.6) a quite loose upper bound to the real VC-dimension. In Figure 2(b), we report similar results for the problem of computing a *relative*ε-close approximation to $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$ for an artificial dataset whose range space has VC-dimension v equal to the length of the longest transaction in the dataset, in this case 33. The dataset contained 100 million transactions. The sample size, given by Lemma 5.2, was computed using $\theta = 0.01$, $\varepsilon = 0.05$, and $\delta = 0.1$. The conclusions we can draw from the results for the behavior of the relative frequency error are similar to those we got for the absolute case. For the case of absolute and relative ε-close approximation to $\mathsf{TOPK}(\mathcal{D}, \mathcal{I}, K)$, we observed results very similar to those obtained for $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$, as expected, given the closed connection between the two problems.

The results of the experiments to evaluate our method to extract a relative ε-close approximation to $\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ are presented in Figure 3(a) and 3(b). The same observations as before hold for the relative frequency error, while it is interesting to note that the relative confidence error is even smaller than the frequency error, most possibly because the confidence of an AR is the ratio between the frequencies of two itemsets that appear in the same transactions, and their sample frequencies will therefore have similar errors that cancel out when the ratio is computed. Similar conclusions can be made for the absolute ε-close approximation case.

From Figures 2(a), 2(b), 3(a), and 3(b), it is also possible to appreciate that, as the sample gets smaller, the maximum and the average errors in the frequency and confidence estimations increase. This suggests that using a fixed sampling rate or a fixed sample size cannot guarantee good results for any ε: not only the estimation of the frequency and/or of the confidence would be quite off from the real value, but because



(a) Relative Association Rule Frequency Error



(b) Relative Association Rule Confidence Error

Fig. 3. Relative ε-close approximation to $\mathsf{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$, artificial dataset, $v = 33$, $\theta = 0.01$, $\gamma = 0.5$, $\varepsilon = 0.05$, $\delta = 0.1$.

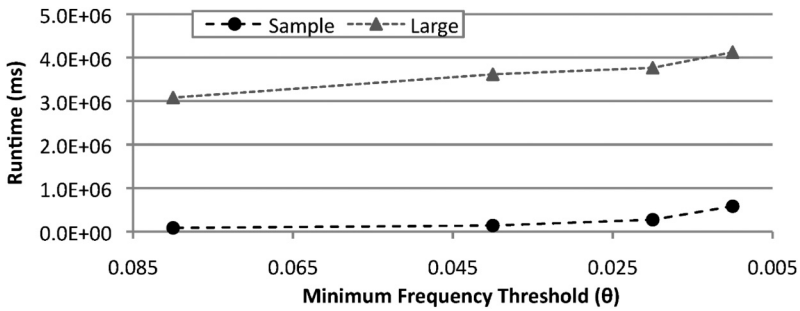
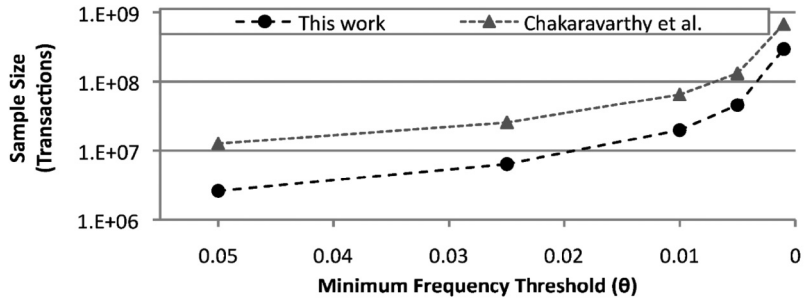


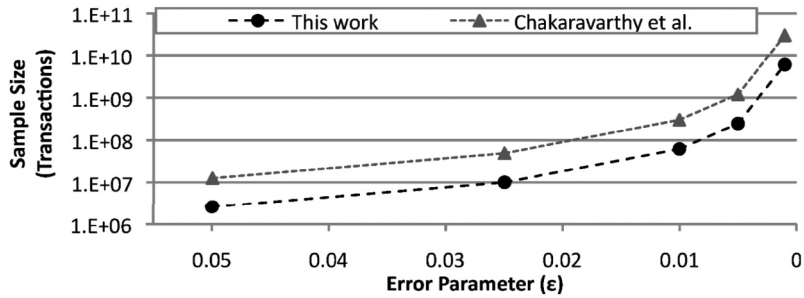
Fig. 4. Runtime Comparison. The sample line includes the sampling time (relative approximation to FIs, artificial dataset, $v = 33$, $\varepsilon = 0.05$, $\delta = 0.1$).

of this, many itemsets that are frequent in the original dataset also may be missing from the output collection and many spurious (very infrequent) itemsets also may be included in it.

The major motivating intuition for the use of sampling in market basket analysis tasks is that mining a sample of the dataset is faster than mining the entire dataset. Nevertheless, the mining time depends not only on the number of transactions, but also on the number of FIs. Given that our methods suggest mining the sample at a lowered minimum frequency threshold, this may cause an increase in running time that would make our method not useful in practice because there may be many more FIs than at the original frequency threshold. We performed a number of experiments to evaluate whether this was the case and present the results in Figure 4. We mined



(a) Sample size as function of θ, $\varepsilon = 0.05$



(b) Sample size as function of ε, $\theta = 0.05$

Fig. 5. Comparison of sample sizes for relative ε-close approximations to $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$. $\Delta = v = 50$, $\delta = 0.05$.

the artificial dataset introduced before for different values of θ and created samples of size sufficient to obtain a relative ε-close approximation to $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$, for $\varepsilon = 0.05$ and $\delta = 0.1$. Figure 4 shows the time needed to mine the large dataset and the time needed to create and mine the samples. It is possible to appreciate that, even considering the sampling time, the speed up achieved by our method is around the order of magnitude (i.e., 10 speed improvement), proving the usefulness of sampling. Moreover, given that the sample size, and therefore the time needed to mine the sample, does not grow with the size of the dataset as long as the d-bound remains constant, that the d-index computation can be performed online, and that the time to create the sample can be made dependent only on the sample size using Vitter's Method D algorithm [Vitter 1987], our method is very scalable as the dataset grows, and the speed-up becomes even more relevant because the mining time for the large dataset would instead grow with the size of the dataset.

Comparing our results to previous work, we note that the bounds generated by our technique are always linear in the VC-dimension v associated with the dataset. As reported in Table I, the best previous work [Chakaravarthy et al. 2009] presented bounds that are linear in the maximum transaction length Δ for two of the six problems studied here. Figures 5(a) and 5(b) show a comparison of the actual sample sizes for relative ε-close approximations to $\mathsf{FI}(\mathcal{D}, \mathcal{I}, \theta)$ as functions of θ and ε. To compute the points for these figures, we set $\Delta = v = 50$, corresponding to the worst possible case for our method; that is, when the VC-dimension of the range space associated to the dataset is exactly equal to the maximum transaction length. We also fixed $\delta = 0.05$ (the two methods behave equally as δ changes). For Figure 5(a), we fixed $\varepsilon = 0.05$, whereas for Figure 5(b) we fixed $\theta = 0.05$. From Figure 5(a) and 5(b), we can appreciate that both bounds have similar but not equal dependencies on θ and ε. More precisely, the bound presented in this work is less dependent on ε and only slightly more dependent on θ.

Table II. Values for Maximum Transaction Length Δ and d-bound q for Real Datasets

	accidents	BMS-POS	BMS-Webview-1	kosarak	pumsb*	retail	webdocs
Δ	51	164	267	2497	63	76	71472
q	46	81	57	443	59	58	2452

It is also evident that the sample sizes given by the bound presented in this work are always much smaller than those presented in Chakaravarthy et al. [2009] (the vertical axis has logarithmic scale). In this comparison, we used $\Delta = v$, but almost all real datasets we encountered have $v \ll \Delta$, as shown in Table II, which would result in a larger gap between the sample sizes provided by the two methods. On the other hand, we should mention that the sample size given by Chakaravarthy et al. [2009] can be slightly optimized by using a stricter version of the Chernoff bound, but this does not change the fact that it depends on the maximum transaction length rather than on the VC-dimension.

7. CONCLUSION

In this article, we presented a novel technique to derive random sample sizes sufficient to easily extract high-quality approximations of the (top-K) FIs and of the collection of ARs. The sample size are linearly dependent on the VC-dimension of the range space associated to the dataset, which is upper bounded by the maximum integer d such that there are at least d transactions of length at least d in the dataset. This bound is tight for a large family of datasets.

We used theoretical tools from statistical learning theory to develop a very practical solution to an important problem in computer science. The practicality of our method is demonstrated in the extensive experimental evaluation that confirmed our theoretical analysis and suggests that, in practice, it is possible to achieve even better results than what the theory guarantees. Moreover, we used this method as the basic building block of an algorithm for the MapReduce [Dean and Ghemawat 2004] distributed/parallel framework of computation. PARMA [Riondato et al. 2012], our MapReduce algorithm, computes an absolute ε-approximation of the collection of FIs or ARs by mining a number of small random samples of the dataset in parallel and then aggregating and filtering the collections of patterns that are frequent in the samples. It allows us to achieve very high-quality approximations of the collection of interest with very high confidence while exploiting and adapting to the available computational resources and achieving a high level of parallelism, highlighted by the quasi-linear speedup we measured while testing PARMA.

Samples of size as computed by our methods can be used to mine approximations of other collection of itemsets, provided that one correctly defines the approximation by taking into account the guarantees on the estimation of the frequency provided by the ε-approximation theorem. For example, one can can use techniques like those presented in Mampaey et al. [2011] on a sample to obtain a small collection of patterns that describe the dataset as well as possible.

We believe that methods and tools developed in the context of computational learning theory can be applied to many problems in data mining and that results traditionally considered of only theoretical interest can be used to obtain very practical methods to solve important problems in knowledge discovery.

It may be possible to develop procedures that give a stricter upper bound to the VC-dimension for a given dataset, or that other measures of sample complexity like the triangular rank [Newman and Rabinovich 2012], shatter coefficients, or Rademacher inequalities [Boucheron et al. 2005], can suggest smaller samples sizes.

ACKNOWLEDGMENTS

The authors are thankful to Luc De Raedt for suggesting the connection between itemsets and monotone monomials. We also thank the anonymous reviewers for their many suggestions and comments for the improvement of this work.

REFERENCES

Ittai Abraham, Daniel Delling, Amos Fiat, Andrew V. Goldberg, and Renato F. Werneck. 2011. VC-dimension and shortest path algorithms. In *Automata, Languages and Programming (Lecture Notes in Computer Science)*, Vol. 6755. Springer, Berlin, 690–699. DOI:http://dx.doi.org/10.1007/978-3-642-22006-7_58

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. *SIGMOD Record* 22 (June 1993), 207–216. Issue 2. DOI:http://dx.doi.org/10.1145/170036.170072

Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487–499.

Noga Alon and Joel H. Spencer. 2008. *The Probabilistic Method* (3rd ed.). John Wiley & Sons, Hoboken, NJ, USA.

Martin Anthony and Peter L. Bartlett. 1999. *Neural Network Learning – Theoretical Foundations*. Cambridge University Press, New York, NY, USA.

Michael Benedikt and Leonid Libkin. 2002. Aggregate operators in constraint query languages. *Journal of Computer System Science* 64, 3 (2002), 628–654. DOI:http://dx.doi.org/DOI: 10.1006/jcss.2001.1810

Avrim Blum, Katrina Ligett, and Aaron Roth. 2008. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC'08)*. ACM, New York, NY, 609–618. DOI:http://dx.doi.org/10.1145/1374376.1374464

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. 1989. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM* 36, 4 (Oct. 1989), 929–965. DOI:http://dx.doi.org/10.1145/76359.76371

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. 2005. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics* 9 (2005), 323–375.

Hervé Brönnimann, Bin Chen, Manoranjan Dash, Peter Haas, and Peter Scheuermann. 2003. Efficient data reduction with EASE. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)*. ACM, New York, NY, 59–68. DOI:http://dx.doi.org/10.1145/956750.956761

Toon Calders, Christophe Rigotti, and Jean-François Boulicaut. 2006. A survey on condensed representations for frequent sets. In *Constraint-Based Mining and Inductive Databases*, Vol. 3848. Springer, Berlin, 64–80. DOI:http://dx.doi.org/10.1007/11615576_4

Aaron Ceglar and John F. Roddick. 2006. Association mining. *ACM Computer Survey* 38, 2 (July 2006), Article 5, 5 pages. Issue 2. DOI:http://dx.doi.org/10.1145/1132956.1132958

Venkatesan T. Chakaravarthy, Vinayaka Pandit, and Yogish Sabharwal. 2009. Analysis of sampling techniques for association rule mining. In *Proceedings of the 12th International Conference on Database Theory (ICDT'09)*. ACM, New York, NY, 276–283. DOI:http://dx.doi.org/10.1145/1514894.1514927

B. Chandra and Shalini Bhaskar. 2011. A new approach for generating efficient sample from market basket data. *Expert Systems with Applications* 38, 3 (2011), 1321–1325. DOI:http://dx.doi.org/DOI: 10.1016/j.eswa.2010.07.008

Bernard Chazelle. 2000. *The Discrepancy Method: Randomness and Complexity*. Cambridge University Press, New York, NY.

Bin Chen, Peter Haas, and Peter Scheuermann. 2002. A new two-phase sampling based algorithm for discovering association rules. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*. ACM, New York, NY, 462–468. DOI:http://dx.doi.org/10.1145/775047.775114

Chyouhwa Chen, Shi-Jinn Horng, and Chin-Pin Huang. 2011. Locality sensitive hashing for sampling-based algorithms in association rule mining. *Expert Systems with Applications* 38, 10 (2011), 12388–12397. DOI:http://dx.doi.org/10.1016/j.eswa.2011.04.018

Yin-Ling Cheung and Ada Wai-Chee Fu. 2004. Mining frequent itemsets without support threshold: With and without item constraints. *IEEE Transactions on Knowledge and Data Engineering* 16, 9 (Sept. 2004), 1052–1069. DOI:http://dx.doi.org/10.1109/TKDE.2004.44

Kun-Ta Chuang, Ming-Syan Chen, and Wen-Chieh Yang. 2005. Progressive sampling for association rules based on sampling error estimation. In *Advances in Knowledge Discovery and Data Mining*, Tu Ho, David Cheung, and Huan Liu (Eds.). Lecture Notes in Computer Science, Vol. 3518. Springer, Berlin, 37–44. DOI:http://dx.doi.org/10.1007/11430919_59

Kun-Ta Chuang, Jiun-Long Huang, and Ming-Syan Chen. 2008. Power-law relationship and self-similarity in the itemset support distribution: analysis and applications. *The VLDB Journal* 17, 5 (Aug. 2008), 1121–1141. DOI:http://dx.doi.org/10.1007/s00778-007-0054-1

Jeffrey Dean and Sanjay Ghemawat. 2004. MapReduce: Simplified data processing on large clusters. In *Proceedings of the 6th Symposium on Operating System Design and Implementation (OSDI'04)*. USENIX Association, 137–150.

Luc Devroye, László Györfi, and Gábor Lugosi. 1996. *A Probabilistic Theory of Pattern Recognition*. Springer, Berlin.

Uriel Feige and Mohammad Mahdian. 2006. Finding small balanced separators. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing (STOC'06)*. ACM, New York, NY, 375–384. DOI:http://dx.doi.org/10.1145/1132516.1132573

Lester R. Ford and Delbert R. Fulkerson. 1962. *Flows in Networks*. Princeton University Press, Princeton, NJ.

Ada Wai-Chee Fu, Renfrew W.-w. Kwong, and Jian Tang. 2000. Mining n-most interesting itemsets. In *Proceedings of the 12th International Symposium on Foundations of Intelligent Systems (ISMIS'00)*. Springer, Berlin, 59–67.

Sorabh Gandhi, Subhash Suri, and Emo Welzl. 2010. Catching elephants with mice: Sparse sampling for monitoring sensor networks. *ACM Transactions on Sensor Networks* 6, 1 (Jan. 2010), Article 1, 27 pages. DOI:http://dx.doi.org/10.1145/1653760.1653761

David Gross-Amblard. 2011. Query-preserving watermarking of relational databases and XML documents. *ACM Transactions on Database Systems* 36, 1 (March 2011), Article 3, 24 pages. DOI:http://dx.doi.org/10.1145/1929934.1929937

Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. 2007. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery* 15, 1 (2007), 55–86. DOI:http://dx.doi.org/10.1007/s10618-006-0059-1

Sariel Har-Peled and Micha Sharir. 2011. Relative (p, ε)-approximations in geometry. *Discrete & Computational Geometry* 45, 3 (2011), 462–496. DOI:http://dx.doi.org/10.1007/s00454-010-9248-1

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, NY.

D. Haussler and E. Welzl. 1986. Epsilon-nets and simplex range queries. In *Proceedings of the 2nd Annual Symposium on Computational Geometry (SCG'86)*. ACM, New York, NY, 61–71. DOI:http://dx.doi.org/10.1145/10515.10522

R. He and J. Shapiro. 2012. Bayesian mixture models for frequent itemsets discovery. *CoRR* abs/1209.6001 (2012), 1–33.

John E. Hopcroft and Richard M. Karp. 1973. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing* 2, 4 (1973), 225–231.

Xuegang Hu and Haitao Yu. 2006. The research of sampling for mining frequent itemsets. In *Rough Sets and Knowledge Technology*, Guo-Ying Wang, James Peters, Andrzej Skowron, and Yiyu Yao (Eds.). Lecture Notes in Computer Science, Vol. 4062. Springer, Berlin, 496–501. DOI:http://dx.doi.org/10.1007/11795131_72

Wontae Hwang and Dongseung Kim. 2006. Improved association rule mining by modified trimming. In *Proceedings of the 6th IEEE International Conference on Computer and Information Technology (CIT'06)*. IEEE Computer Society, 24. DOI:http://dx.doi.org/10.1109/CIT.2006.101

Caiyan Jia and Ruqian Lu. 2005. Sampling ensembles for frequent patterns. In *Fuzzy Systems and Knowledge Discovery*, Lipo Wang and Yaochu Jin (Eds.). Lecture Notes in Computer Science, Vol. 3613. Springer, Berlin, 1197–1206. DOI:http://dx.doi.org/10.1007/11539506_150

Cai-Yan Jia and Xie-Ping Gao. 2005. Multi-scaling sampling: An adaptive sampling method for discovering approximate association rules. *Journal of Computer Science and Technology* 20, 3 (2005), 309–318. DOI:http://dx.doi.org/10.1007/s11390-005-0309-5

George H. John and Pat Langley. 1996. Static versus dynamic sampling for data mining. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)*. The AAAI Press, Menlo Park, CA, 367–370.

J. Kleinberg, M. Sandler, and A. Slivkins. 2008. Network failure detection and graph connectivity. *SIAM Journal on Computing* 38, 4 (2008), 1330–1346. DOI:http://dx.doi.org/10.1137/070697793

Jon M. Kleinberg. 2003. Detecting a network failure. *Internet Mathematics* 1, 1 (2003), 37–55.

Erich L. Lehmann and George Casella. 1998. *Theory of Point Estimation*. Springer-Verlag, New York, NY.

Yanrong Li and Raj Gopalan. 2005. Effective sampling for mining association rules. In *AI 2004: Advances in Artificial Intelligence*, Geoffrey Webb and Xinghuo Yu (Eds.). Lecture Notes in Computer Science, Vol. 3339. Springer, Berlin, 73–75. DOI:http://dx.doi.org/10.1007/978-3-540-30549-1_35

Yi Li, Philip M. Long, and Aravind Srinivasan. 2001. Improved bounds on the sample complexity of learning. *Journal of Computer System Science* 62, 3 (2001), 516–527. DOI:http://dx.doi.org/10.1006/jcss.2000.1741

Nathan Linial, Yishay Mansour, and Ronald L. Rivest. 1991. Results on learnability and the Vapnik-Chervonenkis dimension. *Information and Computation* 90, 1 (1991), 33–49. DOI:http://dx.doi.org/10.1016/0890-5401(91)90058-A

Maarten Löffler and Jeff M. Phillips. 2009. Shape fitting on point sets with probability distributions. In *Algorithms - ESA 2009*, Amos Fiat and Peter Sanders (Eds.). Lecture Notes in Computer Science, Vol. 5757. Springer, Berlin, 313–324. DOI:http://dx.doi.org/10.1007/978-3-642-04128-0_29

Daniel J. McDonald, Cosma Rohilla Shalizi, and Mark Schervish. 2011. Estimated VC dimension for risk bounds. *arXiv preprint*, arXiv:1111.3404 (2011).

Basel A. Mahafzah, Amer F. Al-Badarneh, and Mohammed Z. Zakaria. 2009. A new sampling technique for association rule mining. *Journal of Information Science* 35, 3 (2009), 358–376. DOI:http://dx.doi.org/10.1177/0165551508100382

Michael Mampaey, Nikolaj Tatti, and Jilles Vreeken. 2011. Tell me what I need to know: Succinctly summarizing data with itemsets. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11)*. ACM, New York, NY, 573–581. DOI:http://dx.doi.org/10.1145/2020408.2020499

Heikki Mannila, Hannu Toivonen, and Inkeri Verkamo. 1994. Efficient algorithms for discovering association rules. In *Proceedings of the KDD Workshop*. AAAI Press, Menlo Park, CA, 181–192.

Jiří Matoušek. 2002. *Lectures on Discrete Geometry*. Springer-Verlag, Secaucus, NJ.

Michael Mitzenmacher and Eli Upfal. 2005. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press.

Thomas Natschläger and Michael Schmitt. 1996. Exact VC-dimension of boolean monomials. *Information Processing Letter* 59, 1 (1996), 19–20. DOI:http://dx.doi.org/10.1016/0020-0190(96)00084-1

Ilan Newman and Yuri Rabinovich. 2012. On multiplicative λ-approximations and some geometric applications. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'12)*. SIAM, 51–67.

Srinivasan Parthasarathy. 2002. Efficient progressive sampling for association rules. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)*. IEEE Computer Society, 354–361. DOI:http://dx.doi.org/10.1109/ICDM.2002.1183923

Andrea Pietracaprina, Matteo Riondato, Eli Upfal, and Fabio Vandin. 2010. Mining top-K frequent itemsets through progressive sampling. *Data Mining and Knowledge Discovery* 21, 2 (2010), 310–326.

Andrea Pietracaprina and Fabio Vandin. 2007. Efficient incremental mining of top-k frequent closed itemsets. In *Discovery Science*, Vincent Corruble, Masayuki Takeda, and Einoshin Suzuki (Eds.). Lecture Notes in Computer Science, Vol. 4755. Springer, Berlin, 275–280. DOI:http://dx.doi.org/10.1007/978-3-540-75488-6_29

Matteo Riondato, Mert Akdere, Uğur Çetintemel, Stanley B. Zdonik, and Eli Upfal. 2011. The VC-dimension of SQL queries and selectivity estimation through sampling. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD'11), Part II*, Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis (Eds.), Vol. 6912. Springer, Berlin, 661–676.

Matteo Riondato, Justin A. DeBrabant, Rodrigo Fonseca, and Eli Upfal. 2012. PARMA: A parallel randomized algorithm for association rules mining in MapReduce. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*, Xue-wen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki (Eds.). ACM, New York, NY, 85–94.

Matteo Riondato and Eli Upfal. 2012. Efficient discovery of association rules and frequent itemsets through sampling with tight performance guarantees. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD'12),* Peter A. Flach, Tijl De Bie, and Nello Cristianini (Eds.), Vol. 7523. Springer, Berlin, 25–41.

Tobias Scheffer and Stefan Wrobel. 2002. Finding the most interesting patterns in a database quickly by using sequential sampling. *Journal of Machine Learning Research* 3 (December 2002), 833–862.

Xuhui Shao, Vladimir Cherkassky, and William Li. 1994. Measuring the VC-dimension using optimized experimental design. *Neural Computation* 12, 8 (2000), 1969–1986.

Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. 2004. Selecting the right objective measure for association analysis. *Information Systems* 29, (2004), 293–313.

Hannu Toivonen. 1996. Sampling large databases for association rules. In *Proceedings of the 22th International Conference on Very Large Data Bases (VLDB'96)*. Morgan Kaufmann, San Francisco, CA, 134–145.

Vladimir N. Vapnik. 1999. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY.

Vladimir N. Vapnik and Alexey J. Chervonenkis. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications* 16, 2 (1971), 264–280. DOI:http://dx.doi.org/10.1137/1116025

Vladimir N. Vapnik, Esther Levin, and Yann Le Cun. 1994. Measuring the VC-dimension of a learning machine. *Neural Computation* 6, 5 (1994), 851–876.

Dinkar Vasudevan and Milan Vojonović. 2009. *Ranking through Random Sampling*. MSR-TR-2009-8 8. Microsoft Research.

Jeffrey Scott Vitter. 1987. An efficient algorithm for sequential random sampling. *ACM Transactions on Mathematics Software* 13, 1 (March 1987), 58–67. DOI:http://dx.doi.org/10.1145/23002.23003

Jianyong Wang, Jiawei Han, Ying Lu, and Petre Tzvetkov. 2005b. TFP: An efficient algorithm for mining top-k frequent closed itemsets. *IEEE Transactions on Knowedge and Data Engineering* 17, 5 (May 2005), 652–664. DOI:http://dx.doi.org/10.1109/TKDE.2005.81

Surong Wang, Manoranjan Dash, and Liang-Tien Chia. 2005a. Efficient sampling: Application to image data. In *Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD'05)*. Tu Bao Ho, David Wai-Lok Cheung, and Huan Liu (Eds.), Vol. 3518. Springer, Berlin, 452–463.

Mohammed J. Zaki, S Parthasarathy, Wei Li, and Mitsunori Ogihara. 1997. Evaluation of sampling for data mining of association rules. In *Proceedings of the 7th International Workshop on Research Issues in Data Engineering (RIDE'97)*. IEEE Computer Society, 42–50. DOI:http://dx.doi.org/10.1109/RIDE.1997.583696

Chengqi Zhang, Shichao Zhang, and Geoffrey I. Webb. 2003. Identifying approximate itemsets of interest in large databases. *Applied Intelligence* 18, 1 (2003), 91–104. DOI:http://dx.doi.org/10.1023/A:1020995206763

Yanchang Zhao, Chengqi Zhang, and Shichao Zhang. 2006. Efficient frequent itemsets mining by sampling. In *Proceeding of the 2006 Conference on Advances in Intelligent IT: Active Media Technology 2006*. IOS Press, Amsterdam, The Netherlands, 112–117.

Received February 2013; revised September 2013; accepted December 2013