An Efficient Rigorous Approach for Identifying Statistically Significant Frequent Itemsets[1]

ADAM KIRSCH and MICHAEL MITZENMACHER
Harvard University, Cambridge, MA, USA
ANDREA PIETRACAPRINA and GEPPINO PUCCI
University of Padova, Italy
ELI UPFAL and FABIO VANDIN
Brown University, Providence, RI, USA

As advances in technology allow for the collection, storage, and analysis of vast amounts of data, the task of screening and assessing the significance of discovered patterns is becoming a major challenge in data mining applications. In this work, we address significance in the context of frequent itemset mining. Specifically, we develop a novel methodology to identify a meaningful support threshold s^* for a dataset, such that the number of itemsets with support at least s^* represents a substantial deviation from what would be expected in a random dataset with the same number of transactions and the same individual item frequencies. These itemsets can then be flagged as statistically significant with a small false discovery rate. We present extensive experimental results to substantiate the effectiveness of our methodology.

Categories and Subject Descriptors: H2.8 [**Database Applications**]: Data Mining

General Terms: Algorithms, Measurement

Additional Key Words and Phrases: Frequent itemset mining; statistical significance; multi-hypothesis test; Poisson approximation; False Discovery Rate

Authors' addresses: Adam Kirsch and Michael Mitzenmacher, Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA. Emails: {kirsch,michaelm}@eecs.harvard.edu. Andrea Pietracaprina, Geppino Pucci, Department of Information Engineering, University of Padova, Italy. Emails: {capri,geppo}@dei.unipd.it. Eli Upfal and Fabio Vandin, Computer Science Department, Brown University, Providence, RI, USA. Email: {eli,vandinfa}@cs.brown.edu,.

The work of A. Kirsch and M. Mitzenmacher was supported, in part, by NSF Grant CNS-0721491 and grants from Cisco Systems Inc., Yahoo!, and Google. The work by A. Pietracaprina, G. Pucci, E. Upfal and F. Vandin was supported, in part, by the EC/IST Project 15964 AEOLUS. The work of E. Upfal was also supported, in part, by NSF awards IIS-0325838 and DMI-0600384, and ONR Award N000140610607. This work was done while F. Vandin was affiliated to University of Padova.

[1]A preliminary version of this work was presented in PODS 2009.

1. INTRODUCTION

The discovery of frequent itemsets in transactional datasets is a fundamental primitive that arises in the mining of association rules and in many other scenarios [Han and Kamber 2001; Tan et al. 2006]. In its original formulation, the problem requires that given a dataset $\mathcal{D}$ of transactions over a set of items $\mathcal{I}$ (i.e., a multiset $\mathcal{D}$ with elements in $2^{\mathcal{I}}$), and a support threshold s, all itemsets $X \subseteq \mathcal{I}$ with support at least s in $\mathcal{D}$ (i.e., contained in at least s transactions) be returned. These high-support itemsets are referred to as *frequent itemsets*.

Since the pioneering paper by Agrawal et al. [Agrawal et al. 1993], a vast literature has flourished, addressing variants of the problem, studying foundational issues, and presenting novel algorithmic strategies or clever implementations of known strategies (see, e.g., [Goethals and Zaki 2003; Goethals et al. 2004]), but many problems remain open [Han et al. 2007]. In particular, assessing the significance of the discovered itemsets, or equivalently, flagging statistically significant discoveries with a limited number of false positive outcomes, is still poorly understood and remains one of the most challenging problems in this area.

The classical framework requires that the user decide what is significant by specifying the support threshold s. Unless specific domain knowledge is available, the choice of such a threshold is often arbitrary [Han and Kamber 2001; Tan et al. 2006] and may lead to a large number of spurious discoveries (false positives) that would undermine the success of subsequent analysis.

In this paper, we develop a rigorous and efficient novel approach for identifying frequent itemsets featuring both a global and a pointwise guarantee on their statistical significance. Specifically, we flag as significant a population of itemsets extracted with respect to a certain threshold, if some global characteristics of the population deviate considerably from what would be expected if the dataset were generated randomly with no correlations between items. Also, we make sure that a large fraction of the itemsets belonging to the returned population are individually significant by enforcing a small False Discovery Rate (FDR) [Benjamini and Hochberg 1995] for the population.

1.1 The model

As mentioned above, the significance of a discovery in our framework is assessed based on its deviation from what would be expected in a random dataset in which individual items are placed in transactions independently.

Formally, let $\mathcal{D}$ be a dataset of t transactions on a set $\mathcal{I}$ of n items, where each transaction is a subset of $\mathcal{I}$. Let $n(i)$ be the number of transactions that contain item i and let $f_i = n(i)/t$ be the *frequency* of item i in the dataset. The *support* of an itemset $X \subseteq \mathcal{I}$ is defined as the number of transactions that contain X. Following [Silverstein et al. 1998], the dataset $\mathcal{D}$ is associated with a probability space of datasets, all featuring the same number of transactions t on the same set of items $\mathcal{I}$ as $\mathcal{D}$, and in which item i is included in any given transaction with probability f_i, independently of all other items and all other transactions. A similar model is used in [Purdom et al. 2004] and [Sayrafi et al. 2005] to evaluate

the running time of algorithms for mining frequent itemsets. For a fixed integer $k \geq 1$, among all possible $\binom{n}{k}$ itemsets of size k (*k-itemsets*) we are interested in identifying statistically significant ones, that is, those k-itemsets whose supports are significantly higher, in a statistical sense, than their expected supports in a dataset drawn at random from the aforementioned probability space.

An alternative probability space of datasets, proposed in [Gionis et al. 2006], considers all arrangements of n items into t transactions which match the exact item frequencies and transaction lengths as in a given dataset $\mathcal{D}$. Conceivably, the technique of this paper could be adapted to this latter model as well.

1.2 Multi-hypothesis testing

In a simple statistical test, a null hypothesis H_0 is tested against an alternative hypothesis H_1. A test consists of a rejection (critical) region C such that, if the statistic (outcome) of the experiment is in C, then the null hypothesis is rejected, and otherwise the null hypothesis is not rejected. The *significance level* of a test, $\alpha = \mathbf{Pr}(\text{Type I error})$, is the probability of rejecting H_0 when it is true (false positive). The *power* of the test, $1-\mathbf{Pr}(\text{Type II error})$, is the probability of correctly rejecting the null hypothesis. The *p-value* of a test is the probability of obtaining an outcome at least as extreme as the one that was actually observed, under the assumption that H_0 is true.

In a multi-hypothesis statistical test, the outcome of an experiment is used to test simultaneously a number of hypotheses. For example, in the context of frequent itemsets, if we seek significant k-itemsets, we are in principle testing $\binom{n}{k}$ null hypotheses simultaneously, where each null hypothesis corresponds to the support of a given itemset not being statistically significant. In the context of multi-hypothesis testing, the significance level cannot be assessed by considering each individual hypothesis in isolation. To demonstrate the importance of correcting for multiplicity of hypotheses, consider a simple real dataset of 1,000,000 transactions over 1,000 items, each with frequency 1/1000. Assume that we observed that a pair of items (i, j) appears in at least 7 transactions. Is the support of this pair statistically significant? To evaluate the significance of this discovery we consider a random dataset where each item is included in each transaction with probability 1/1000, independent of all items. The probability that the pair (i, j) is included in a given transaction is 1/1,000,000, thus the expected number of transactions that include this pair is 1. A simple calculation shows that the probability that (i, j) appears in at least 7 transactions is about 0.0001. Thus, it seems that the support of (i, j) in the real dataset is statistically significant. However, each of the 499,500 pairs of items has probability 0.0001 to appear in at least 7 transactions in the random dataset. Thus, even under the assumption that items are placed independently in transactions, the expected number of pairs with support at least 7 is about 50. If there were only about 50 pairs with support at least 7, returning the pair (i, j) as a statistically significant itemset would likely be a false discovery since its frequency would be better explained by random fluctuations in observed data. On the other hand, assume that the real dataset contains 300 disjoint pairs each with

support at least 7. The probability of that event in the random dataset is less than $\binom{\binom{1000}{2}}{300}10^{-4\times 300} \leq 2^{-300}$. Thus, it is very likely that the support of most of these pairs would be statistically significant. A discovery process that does not return these pairs will result in a large number of false negative errors. Our goal is to design a rigorous methodology which is able to distinguish between these two scenarios.

A natural generalization of the significance level to multi-hypothesis testing is the *Family Wise Error Rate (FWER)*, which is the probability of incurring at least one Type I error in any of the individual tests. If we have m simultaneous tests and we want to bound the FWER by α, then the Bonferroni method tests each null hypothesis with significance level α/m. While controlling the FWER, this method is too conservative in that the power of the test is too low, giving many false negatives. There are a number of techniques that improve on the Bonferroni method, but for large numbers of hypotheses all of these techniques lead to tests with low power (see [Dudoit et al. 2003] for a good review).

The *False Discovery Rate (FDR)* was suggested by Benjamini and Hochberg [Benjamini and Hochberg 1995] as an alternative, less conservative approach to control errors in multiple tests. Let V be the number of Type I errors in the individual tests, and let R be the total number of null hypotheses rejected by the multiple test. Then we define FDR to be the expected ratio of erroneous rejections among all rejections, namely FDR $= E[V/R]$, with $V/R = 0$ when $R = 0$. Designing a statistical test that controls for FDR is not simple, since the FDR is a function of two random variables that depend both on the set of null hypotheses and the set of alternative hypotheses. Building on the work of [Benjamini and Hochberg 1995], Benjamini and Yekutieli [Benjamini and Yekutieli 2001] developed a general technique for controlling the FDR in any multi-hypothesis test (see Theorem 5 in Section 3.1).

1.3 Our Results

We address the classical problem of mining frequent itemsets with respect to a certain minimum support threshold, and provide a rigorous methodology to establish a threshold that guarantees, in a statistical sense, that the returned family of frequent itemsets contains significant ones with a limited FDR. Our methodology crucially relies on the following Poisson approximation result, which is the main theoretical contribution of the paper.

Consider a dataset $\mathcal{D}$ of t transactions on a set $\mathcal{I}$ of n items and let $\hat{\mathcal{D}}$ be random dataset from the probability space associated with $\mathcal{D}$ as described in Section 1.1. Let $Q_{k,s}$ be the observed number of k-itemsets with support at least s in $\mathcal{D}$, and let $\hat{Q}_{k,s}$ be the corresponding random variable for $\hat{\mathcal{D}}$. We show that there exists a minimum support value $s_{\min}$ (which depends on the parameters of $\mathcal{D}$ and on k), such that for all $s \geq s_{\min}$ the distribution of $\hat{Q}_{k,s}$ is well approximated by a Poisson distribution. Our result is based on a novel application of the Chen-Stein Poisson approximation method [Arratia et al. 1990].

The minimum support $s_{\min}$ provides the grounds to devise a rigorous method for

establishing a support threshold for mining significant itemsets, both reducing the overall complexity and improving the accuracy of the discovery process. Specifically, for a fixed itemset size k, we test a small number of support thresholds $s \geq s_{\min}$, and, for each such threshold, we measure the p-value corresponding to the null hypothesis H_0 that the observed value $Q_{k,s}$ comes from a Poisson distribution of suitable expectation. From the tests we can determine a threshold s^* such that, with user-defined significance level α, the number of k-itemsets with support at least s^* is not sampled from a Poisson distribution and is therefore statistically significant. Observe that the statistical significance of the number of itemsets with support at least s^* does not imply necessarily that each of the itemsets is significant. However, our test is also able to guarantee a user-defined upper bound β on the FDR among all discoveries. We remark that our approach works for any fixed itemset size k, unlike traditional frequent itemset mining, where itemsets of all sizes are extracted for a given threshold.

To grasp the intuition behind the above approach, recall that a Poisson distribution models the number of occurrences among a large set of possible events, where the probability of each event is small. In the context of frequent itemset mining, the Poisson approximation holds when the probability that an individual itemset has support at least $s_{\min}$ in $\hat{\mathcal{D}}$ is small, and thus the existence of such an event in $\mathcal{D}$ is likely to be statistically significant. We stress that our technique discovers statistically significant itemsets among those of relatively high support. In fact, if the expected supports of individual itemsets vary in a large range, there may exist itemsets with very low expected supports in $\hat{\mathcal{D}}$ which may have statistically significant supports in $\mathcal{D}$. These itemsets would not be discovered by our strategy. However, any mining strategy aiming at discovering significant, low-support itemsets is likely to incur high costs due to the large (possibly exponential) number of candidates to be examined, although only a few of them would turn out to be significant.

We validate our theoretical results by mining significant frequent itemsets from a number of real datasets that are standard benchmarks in this field. Also, we compare the performance of our methodology to a standard multi-hypothesis approach based on [Benjamini and Yekutieli 2001], and provide evidence that the latter often returns fewer significant itemsets, which indicates that our method has considerably higher power.

1.4 Related Work

A number of works have explored various notions of significant itemsets and have proposed methods for their discovery. Below, we review those most relevant to our approach and refer the reader to [Han et al. 2007, Section 3] for further references. Aggarwal and Yu [Aggarwal and Yu 1998] relate the significance of an itemset X to the quantity $((1 - v(X))/(1 - \mathbf{E}[v(X)])) \cdot (\mathbf{E}[v(X)]/v(X))$, where $v(X)$ represents the fraction of transactions containing some but not all of the items of X, and $\mathbf{E}[v(X)]$ represents the expectation of $v(X)$ in a random dataset where items occur in transactions independently. This ratio provides an empirical measure of

the correlation among the items of X that, according to the authors, is more effective than absolute support. In [Srikant and Agrawal 1996; DuMouchel 1999; DuMouchel and Pregibon 2001], the significance of an itemset is measured as the ratio R between its actual support and its expected support in a random dataset. In order to make this measure more accurate for small supports, [DuMouchel 1999; DuMouchel and Pregibon 2001] propose smoothing the ratio R using an empirical Bayesian approach. Bayesian analysis is also employed in [Silberschatz and Tuzhilin 1996] to derive subjective measures of significance of patterns (e.g., itemsets) based on how strongly they "shake" a system of established beliefs. In [Jaroszewicz and Scheffer 2005], the significance of an itemset is defined as the absolute difference between the support of the itemset in the dataset, and the estimate of this support made from a Bayesian network with parameters derived from the dataset.

A statistical approach for identifying significant itemsets is presented in [Silverstein et al. 1998], where the measure of interest for an itemset is defined as the degree of dependence among its constituent items, which is assessed through a χ^2 test. Unfortunately, as reported in [DuMouchel 1999; DuMouchel and Pregibon 2001], there are technical flaws in the applications of the statistical test in [Silverstein et al. 1998]. Nevertheless, this work pioneered the quest for a rigorous framework for addressing the discovery of significant itemsets.

A common drawback of the aforementioned works is that they assess the significance of each itemset *in isolation*, rather than taking into account the *global* characteristics of the dataset from which they are extracted. As argued before, if the number of itemsets considered by the analysis is large, even in a purely random dataset some of them are likely to be flagged as significant if considered in isolation. A few works attempt at accounting for the global structure of the dataset in the context of frequent itemset mining. The authors of [Gionis et al. 2006] propose an approach based on Markov chains to generate a random dataset that has identical transaction lengths and identical frequencies of the individual items as the given real dataset. The work suggests comparing the outcomes of a number of data mining tasks, frequent itemset mining among the others, in the real and the randomly generated datasets in order to assess whether the real datasets embody any significant global structure. However, such an assessment is carried out in a purely qualitative fashion without rigorous statistical grounding.

The problem of spurious discoveries in the mining of significant patterns is studied in [Bolton et al. 2002]. The paper is concerned with the discovery of significant pairs of items, where significance is measured through the p-value, that is, the probability of occurrence of the observed support in a random dataset. Significant pairs are those whose p-values are below a certain threshold that can be suitably chosen to bound the FWER, or to bound the FDR. The authors compare the relative power of the two metrics through experimental results, but do not provide methods to set a meaningful support threshold, which is the most prominent feature of our approach.

Beyond frequent itemset mining, the issue of significance has also been addressed in the realm of discovering association rules. In [Hämäläinen and Nykänen 2008],

Dataset	n	$[f_{\min}; f_{\max}]$	m	t
Retail	16470	[1.13e-05 ; 0.57]	10.3	88162
Kosarak	41270	[1.01e-06 ; 0.61]	8.1	990002
Bms1	497	[1.68e-05 ; 0.06]	2.5	59602
Bms2	3340	[1.29e-05 ; 0.05]	5.6	77512
Bmspos	1657	[1.94e-06 ; 0.60]	7.5	515597
Pumsb*	2088	[2.04e-05 ; 0.79]	50.5	49046

Table I. Parameters of the benchmark datasets: $\boldsymbol{n}$ is the number of items; $[\boldsymbol{f_{\min}}, \boldsymbol{f_{\max}}]$ is the range of frequencies of the individual items; $\boldsymbol{m}$ is the average transaction length; and $\boldsymbol{t}$ is the number of transactions.

the authors provide a variation of the well-known Apriori strategy for the efficient discovery of a subset $\mathcal{A}$ of association rules with p-value below a given cutoff value, while the results in [Megiddo and Srikant 1998] provide the means of evaluating the FDR in $\mathcal{A}$. The FDR metric is also employed in [Zhang et al. 2004] for the discovery of significant quantitative rules, a variation of association rules. None of these works is able to establish support thresholds such that the returned discoveries feature small FDR.

1.5 Benchmark datasets

In order to validate the methodology, a number of experiments, whose results are reported in Section 4, have been performed on datasets which are standard benchmarks in the context of frequent itemsets mining. The main characteristics of the datasets we use are summarized in Table I. A description of the datasets can be found in the FIMI Repository (`http://fimi.cs.helsinki.fi/data/`), where they are available for download.

1.6 Organization of the Paper

The rest of the paper is structured as follows. Section 2 presents the Poisson approximation result for the random variable $\hat{Q}_{k,s}$. The methodology for establishing the support threshold s^* is presented in Section 3, and experimental results are reported in Section 4. Section 5 ends the paper with some concluding remarks.

2. POISSON APPROXIMATION RESULT

The Chen-Stein method [Arratia et al. 1990] is a powerful tool for bounding the error in approximating probabilities associated with a sequence of dependent events by a Poisson distribution. To apply the method to our case, we fix parameters k and s, and define a collection of $\binom{n}{k}$ Bernoulli random variables $\{Z_{X,s} \mid X \subset \mathcal{I},\ |X| = k\}$, such that $Z_{X,s} = 1$ if the k-itemset X appears in at least s transactions in the random dataset $\hat{\mathcal{D}}$, and $Z_{X,s} = 0$ otherwise. Also, let $p_X = \mathbf{Pr}(Z_{X,s} = 1)$. We are interested in the distribution of $\hat{Q}_{k,s} = \sum_{X:|X|=k} Z_{X,s}$.

For each set X we define the *neighborhood set* of X,

$$I(X) = \{X' \mid X \cap X' \neq \emptyset, |X'| = |X|\}.$$

If $Y \notin I(X)$ then $Z_{Y,s}$ and $Z_{X,s}$ are independent. The following theorem is a

straightforward adaptation of [Arratia et al. 1990, Theorem 1] to our case.

THEOREM 1. *Let U be a Poisson random variable such that $\mathbf{E}[U] = \mathbf{E}[\hat{Q}_{k,s}] = \lambda < \infty$. The variation distance between the distributions $\mathcal{L}(\hat{Q}_{k,s})$ of $\hat{Q}_{k,s}$ and $\mathcal{L}(U)$ of U is such that*

$$\left\|\mathcal{L}(\hat{Q}_{k,s}) - \mathcal{L}(U)\right\| = \sup_A |\mathbf{Pr}(\hat{Q}_{k,s} \in A) - \mathbf{Pr}(U \in A)| \leq b_1 + b_2,$$

where

$$b_1 = \sum_{X:|X|=k} \sum_{Y \in I(X)} p_X p_Y$$

and

$$b_2 = \sum_{X:|X|=k} \sum_{X \neq Y \in I(X)} \mathbf{E}[Z_{X,s} Z_{Y,s}].$$

We can derive analytic bounds for b_1 and b_2 in many situations. Specifically, suppose that we generate t transactions in the following way. For each item x, we sample a random variable $R_x \in [0,1]$ independently from some distribution R. Conditioned on the R_x's, each item x occurs independently in each transaction with probability R_x. In what follows, we provide specific bounds for this situation that depend on the moment $\mathbf{E}[R^{2s}]$ of the random variable R.

As a warm-up, we first consider the specific case where each R_x is a fixed value $p = \gamma/n$ for some constant γ for all x. That is, each item appears in each transaction with a fixed probability p, and the expected number of items per transaction is constant. The more general case follows the same approach, albeit with a few more technical difficulties.

THEOREM 2. *Consider an asymptotic regime where as $n \to \infty$, we have $k, s = O(1)$ with $s \geq 2$, each item appears in each transaction with probability $p = \gamma/n$ for some constant γ, and $t = O(n^c)$ for some positive constant $0 < c \leq (k-1)(1-1/s)$. Let U be a Poisson random variable such that $\mathbf{E}[U] = \mathbf{E}[\hat{Q}_{k,s}] = \lambda < \infty$. Then the variation distance between the distributions $\mathcal{L}(\hat{Q}_{k,s})$ of $\hat{Q}_{k,s}$ and $\mathcal{L}(U)$ of U satisfies*

$$\left\|\mathcal{L}(\hat{Q}_{k,s}) - \mathcal{L}(U)\right\| = O(1/n^{2s-2}).$$

PROOF. For a given set X of k items, let $p_{X,i}$ be the probability that X appears in exactly i transactions, so that $p_X = \sum_{i=s}^{t} p_{X,i}$ and

$$p_{X,i} = \binom{t}{i} \left(\frac{\gamma}{n}\right)^{ki} \left(1 - \left(\frac{\gamma}{n}\right)^k\right)^{t-i}.$$

Applying Theorem 1 gives

$$\left\|\mathcal{L}(\hat{Q}_{k,s}) - \mathcal{L}(U)\right\| \leq b_1 + b_2$$

where

$$b_1 = \sum_{X:|X|=k} \sum_{Y \in I(S)} p_X p_Y$$

and

$$b_2 = \sum_{X:|S|=k} \sum_{Y \neq X \in I(S)} \mathbf{E}[Z_{X,s} Z_{Y,s}].$$

We now evaluate b_1 and b_2. A direct calculation easily gives the value for b_1 given in the statement of the theorem. For the asymptotic analysis, we write

$$\begin{aligned}
&\left(\binom{n}{k}^2 - \binom{n}{k}\binom{n-k}{k}\right) \\
&= \binom{n}{k}^2 \left(1 - \frac{\binom{n-k}{k}}{\binom{n}{k}}\right) \\
&= \binom{n}{k}^2 \left(1 - \prod_{i=0}^{k-1} \frac{n-k-i}{n-i}\right) \\
&= \Theta(n^k)^2 \cdot \Theta(1/n) = \Theta(n^{2k-1})
\end{aligned}$$

and

$$\begin{aligned}
p_{X,s} &= \binom{t}{s} \left(\frac{\gamma}{n}\right)^{ks} \left(1 - \left(\frac{\gamma}{n}\right)^k\right)^{t-s} \\
&= \Theta(t^s) \cdot \Theta(n^{-ks}) \cdot (1 + o(1)) = \Theta\left(t^s n^{-ks}\right),
\end{aligned}$$

where we have used the fact that $t = o(n^k)$ to obtain the asymptotics for the third term. Also, we note that for any $1 \leq i < t$

$$\frac{p_{X,i+1}}{p_{X,i}} = \frac{t-i}{i+1} \left(\frac{\gamma}{n}\right)^k \left(1 - \left(\frac{\gamma}{n}\right)^k\right)^{-1}$$

and so

$$\max_{i \in \{s, s+1, \ldots, t-1\}} \frac{p_{X,i+1}}{p_{X,i}} = O(tn^{-k}) = O(1/n).$$

Using a geometric series, it follows that

$$p_X = \sum_{i=s}^{t} p_{X,i} = p_{X,s}(1 + o(1)) = \Theta\left(t^s n^{-ks}\right).$$

Thus, we obtain

$$\begin{aligned}
b_1 &= \Theta(n^{2k-1}) \cdot \Theta\left(t^s n^{-ks}\right)^2 \\
&= \Theta(t^{2s} n^{2k(1-s)-1}) = \Theta(n^{2cs+2k(1-s)-1}).
\end{aligned}$$

We now turn our attention to b_2. Consider sets $X \neq Y$ of k items, let $g = |X \cap Y|$, and suppose that $g > 0$. Then if $Z_{X,s} Z_{Y,s} = 1$, there exist disjoint subsets $A, B, C \in \{1, \ldots, t\}$ such that $0 \leq |A| \leq s$, $|B| = |C| = s - |A|$, all of the transactions in A contain both X and Y, all of the transactions in B contain X, and all of the transactions in C contain Y.

Therefore,

$$\mathbf{E}[Z_{X,s}Z_{Y,s}] \leq \sum_{i=0}^{s} \binom{t}{i; s-i; s-i} \left(\frac{\gamma}{n}\right)^{(2k-g)i+2k(s-i)},$$

where the notation $\binom{m}{x;y;z}$ is a shorthand for $\binom{m}{x}\binom{m-x}{y}\binom{m-x-y}{z}$.

It follows that

$$\begin{aligned}
b_2 &\leq \sum_{g=1}^{k-1} \binom{n}{g; k-g; k-g} \\
&\quad \times \sum_{i=0}^{s} \binom{t}{i; s-i; s-i} \left(\frac{\gamma}{n}\right)^{(2k-g)i+2k(s-i)} \\
&= \sum_{g=1}^{k-1} \binom{n}{g; k-g; k-g} \left(\frac{\gamma}{n}\right)^{2ks} \\
&\quad \times \sum_{i=0}^{s} \binom{t}{i; s-i; s-i} \left(\frac{n}{\gamma}\right)^{gi} \\
&= \sum_{g=1}^{k-1} \binom{n}{g; k-g; k-g} \left(\frac{\gamma}{n}\right)^{2ks} \\
&\quad \times \sum_{i=0}^{s} \binom{t}{i; s-i; s-i} \left(\frac{n}{\gamma}\right)^{gi} \\
&= \sum_{g=1}^{k-1} \Theta(n^{2k-g+2cs}) \left(\frac{\gamma}{n}\right)^{2ks} \sum_{i=0}^{s} n^{-ic} \left(\frac{n}{\gamma}\right)^{gi} \\
&= \Theta(n^{2k(1-s)+2cs}) \sum_{g=1}^{k-1} n^{-g} \sum_{i=0}^{s} \gamma^{-gi} n^{(g-c)i} \\
&= \Theta(n^{2k(1-s)+2cs}) \sum_{g=1}^{k-1} n^{-g} \begin{cases} \Theta(1) & g \leq c \\ \Theta(n^{(g-c)s}) & g > c \end{cases} \\
&= \Theta(n^{2k(1-s)+2cs}) \cdot \Theta(n^{-(k-1)+(k-1-c)s}) \\
&= \Theta(n^{2k(1-s)+s(k-1+c)-k+1})
\end{aligned}$$

Note that, in the summation where there are two cases depending on whether $g \leq c$ or $g > c$, we have used the assumption that $c \leq (k-1)(1-1/s)$ to ensure the next equality. Finally, it is simple to check that both b_1 and b_2 are $O(1/n^{2s-2})$ if $c \leq (k-1)(1-1/s)$. □

We now provide the more general theorem.

THEOREM 3. *Consider an asymptotic regime where as $n \to \infty$, we have $k, s = O(1)$ with $s \geq 2$, $\mathbf{E}[R^{2s}] = O(n^{-a})$ for some constant $2 < a \leq 2s$, and $t = O(n^c)$ for some positive constant c. Let U be a Poisson random variable such that $\mathbf{E}[U] =$*

$\mathbf{E}[\hat{Q}_{k,s}] = \lambda < \infty$. *If*

$$c \leq \frac{(k-1)(a-2) + \min(2a-6, 0)}{2s},$$

then the variation distance between the distributions $\mathcal{L}(\hat{Q}_{k,s})$ *of* $\hat{Q}_{k,s}$ *and* $\mathcal{L}(U)$ *of* U *satisfies*

$$\left\| \mathcal{L}(\hat{Q}_{k,s}) - \mathcal{L}(U) \right\| = O(1/n).$$

PROOF. Applying Theorem 1 gives

$$\left\| \mathcal{L}(\hat{Q}_{k,s}) - \mathcal{L}(U) \right\| \leq b_1 + b_2$$

where

$$b_1 = \sum_{X:|X|=k} \sum_{Y \in I(X)} p_X p_Y$$

and

$$b_2 = \sum_{X:|X|=k} \sum_{Y \neq X \in I(X)} \mathbf{E}[Z_{X,s} Z_{Y,s}].$$

We now evaluate b_1 and b_2. Letting $\vec{R}$ denote the vector of the R_x's, we have that for any set X of k items

$$\mathbf{Pr}(Z_{X,s} = 1 \mid \vec{R}) \leq \binom{t}{s} \prod_{x \in X} R_x^s.$$

Since the R_x's are independent with common distribution R,

$$p_X = \mathbf{E}[\mathbf{Pr}(Z_{X,s} = 1 \mid \vec{R})] \leq \binom{t}{s} \mathbf{E}[R^s]^k.$$

Using Jensen's inequality, we now have

$$\begin{aligned}
b_1 &= \sum_{X:|X|=k} \sum_{Y \in I(X)} p_X p_Y \\
&\leq \left(\binom{n}{k}^2 - \binom{n}{k}\binom{n-k}{k} \right) \binom{t}{s}^2 \mathbf{E}[R^s]^{2k} \\
&\leq \binom{n}{k}^2 \left(1 - \frac{\binom{n-k}{k}}{\binom{n}{k}} \right) \binom{t}{s}^2 \mathbf{E}[R^{2s}]^k \\
&= \binom{n}{k}^2 \left(1 - \prod_{i=0}^{k-1} \frac{n-k-i}{n-i} \right) \binom{t}{s}^2 \mathbf{E}[R^{2s}]^k \\
&= \Theta(n^k)^2 \cdot \Theta(1/n) \cdot O(n^{2cs}) \cdot O(n^{-ka}) \\
&= O(n^{k(2-a)+2cs-1})
\end{aligned}$$

We now turn our attention to b_2. Consider sets $X \neq Y$ of k items, and suppose $g = |X \cap Y| > 0$. If $Z_{X,s}Z_{Y,s} = 1$, there exist disjoint subsets $A, B, C \in \{1, \ldots, t\}$ such that $0 \leq |A| \leq s$, $|B| = |C| = s - |A|$, all of the transactions in A contain both X and Y, all of the transactions in B contain X, and all of the transactions in C contain Y. Therefore,

$$\begin{aligned}
\mathbf{E}[Z_{X,s}Z_{Y,s} \mid \vec{R}] &\leq \sum_{i=0}^{s} \binom{t}{i; s-i; s-i} \left(\prod_{x \in X \cup Y} R_x^i\right) \\
&\qquad \times \left(\prod_{x \in X} R_x^{s-i}\right)\left(\prod_{y \in Y} R_y^{s-i}\right) \\
&= \sum_{i=0}^{s} \binom{t}{i; s-i; s-i} \left(\prod_{x \in X \cap Y} R_x^{2s-i}\right) \\
&\qquad \times \left(\prod_{x \in X - Y} R_x^{s}\right)\left(\prod_{y \in Y - X} R_y^{s}\right).
\end{aligned}$$

Applying independence of the R_x's and Jensen's inequality gives

$$\begin{aligned}
\mathbf{E}[Z_{X,s}Z_{Y,s}] &= \mathbf{E}[\mathbf{E}[Z_{X,s}Z_{Y,s} \mid \vec{R}]] \\
&\leq \sum_{i=0}^{s} \binom{t}{i; s-i; s-i} \mathbf{E}[R^{2s-i}]^g \mathbf{E}[R^s]^{2(k-g)} \\
&\leq \sum_{i=0}^{s} t^{2s-i} \mathbf{E}[R^{2s}]^{\frac{g(2s-i)}{2s}} \mathbf{E}[R^{2s}]^{k-g} \\
&= \sum_{i=0}^{s} t^{2s-i} \mathbf{E}[R^{2s}]^{k-ig/2s} \\
&\leq O(1) \sum_{i=0}^{s} n^{(2s-i)c - a(k-ig/2s)} \\
&= O(n^{2sc-ak}) \sum_{i=0}^{s} n^{i\left(\frac{ag}{2s} - c\right)} \\
&= O\left(n^{2sc-ak+\max\left\{0, s\left(\frac{ag}{2s}-c\right)\right\}}\right)
\end{aligned}$$

It follows that

$$\begin{aligned}
b_2 &\leq \sum_{g=1}^{k-1} \binom{n}{g; k-g; k-g} O\left(n^{2sc-ak+\max\left\{0, s\left(\frac{ag}{2s}-c\right)\right\}}\right) \\
&= O(n^{2k+2sc-ak}) \sum_{g=1}^{k-1} n^{-g} O\left(n^{\max\left\{0, s\left(\frac{ag}{2s}-c\right)\right\}}\right)
\end{aligned}$$

Now, for $2sc/a < g < k$, we have (using the fact that $a \geq 2$)

$$n^{-g} n^{\max\{0, s(\frac{ag}{2s} - c)\}} = n^{g(\frac{a}{2}-1)-sc} \leq n^{(k-1)(\frac{a}{2}-1)-sc}.$$

Thus

$$b_2 = O(n^{2k+sc-ak+(k-1)(\frac{a}{2}-1)}).$$

(Here we are using the fact that our choice of c satisfies $c \leq (k-1)(a-2)/2s$ to ensure that $n^{(k-1)(\frac{a}{2}-1)-cs} = \Omega(1)$.)

Now, we have

$$b_1 = O(1/n)$$

since

$$c \leq \frac{(k-1)(a-2)}{2s} \leq \frac{k(a-2)}{2s},$$

and

$$b_2 = O(1/n)$$

since

$$c \leq \frac{k(a-2)+(a-4)}{2s}.$$

Thus

$$b_1 + b_2 = O(1/n).$$

□

It is easy to see that for fixed k, the quantities b_1 and b_2 defined in Theorem 1 are both decreasing in s. In the following, we will use the notation $b_1(s)$ and $b_2(s)$ to indicate explicitly that both quantities are functions of s. Therefore, for a chosen ϵ, with $0 < \epsilon < 1$, we can define

$$s_{\min} = \min\{s \geq 1 \ : \ b_1(s) + b_2(s) \leq \epsilon\}. \quad (1)$$

It immediately follows that for every s in the range $[s_{\min}, \infty)$, the variation distance between the distribution of $\hat{Q}_{k,s}$ and the distribution of a Poisson variable with the same expectation is less than ϵ. In other words, for every $s \geq s_{\min}$ the number of k-itemsets with support at least s is well approximated by a Poisson variable. Theorems 2 and 3 proved above establish the existence of meaningful ranges of s for which the Poisson approximation holds, under certain constraints on the individual item frequencies in the random dataset and on the other parameters.

2.1 A Monte Carlo method for determining $s_{\min}$

While the analytical results of the previous subsection require that the individual item frequencies in the random dataset be drawn from a given distribution, in what follows we give experimental evidence that the Poisson approximation for the distribution of $\hat{Q}_{k,s}$ holds also when the item frequencies are fixed arbitrarily, as is the case of our reference random model. More specifically, we present a method

which approximates the support threshold $s_{\min}$ defined by Equation 1, based on a simple Monte Carlo simulation which returns estimates of $b_1(s)$ and $b_2(s)$. This approach is also convenient in practice since it avoids the inevitable slack due to the use of asymptotics in Theorem 3.

For a given configuration of item frequencies and number of transactions, let $\tilde{s}$ be the maximum expected support of any k-itemset in a random dataset sampled according to that configuration, that is, the product of the k largest item frequencies. Conceivably, the value $b_1(\tilde{s})$ is rather large, hence it makes sense to search for an $s_{\min}$ larger than $\tilde{s}$. We generate Δ random datasets and from each such dataset we mine all of the k-itemsets of support at least $\tilde{s}$. Let W be the set of itemsets extracted in this fashion from all of the generated datasets. For each $s \geq \tilde{s}$ we can estimate $b_1(s)$ and $b_2(s)$ by computing for each $X \in W$ the empirical probability p_X of the event $Z_{X,s} = 1$, and for each pair $X, Y \in W$, with $X \cap Y \neq \emptyset$, the empirical probability $p_{X,Y}$ of the event $Z_{X,s}Z_{Y,s} = 1$. Note that for itemsets not in W these probabilities are estimated as 0. If it turns out that $b_1(\tilde{s}) + b_2(\tilde{s}) > \epsilon/4$, then we let $\hat{s}_{\min}$ be the minimum $s > \tilde{s}$ such that $b_1(s) + b_2(s) \leq \epsilon/4$. Otherwise, if $b_1(\tilde{s}) + b_2(\tilde{s}) \leq \epsilon/4$, we repeat the above procedure starting from $\tilde{s}/2$. (Based on the above considerations this latter case will be unlikely.) Algorithm 1 implements the above ideas.

The following theorem provides a bound on the probability that $\hat{s}_{\min}$ be a conservative estimate of $s_{\min}$, that is, $\hat{s}_{\min} \geq s_{\min}$.

THEOREM 4. *If $\Delta = O\left(\log(1/\delta)/\epsilon\right)$, the output $\hat{s}_{\min}$ of the Monte-Carlo process satisfies*

$$\mathbf{Pr}(b_1(\hat{s}_{\min}) + b_2(\hat{s}_{\min}) \leq \epsilon) \geq 1 - \delta.$$

PROOF. Let assume $b_1(\hat{s}_{\min}) + b_2(\hat{s}_{\min}) > \epsilon$. Note that $b_1(\hat{s}_{\min}) \leq b_2(\hat{s}_{\min})$, therefore we have $b_2(\hat{s}_{\min}) > \epsilon/2$. Let B be the random variable corresponding to Δ times the estimate of $b_2(\hat{s}_{\min})$ obtained with Algorithm 1. Thus $E[B] > \Delta\epsilon/2$. Since Algorithm 1 returns $\hat{s}_{\min}$ as estimate of $s_{\min}$, we have that $B \leq \Delta\epsilon/4$. Let

$$\Delta = \frac{8\log(1/\delta)}{\epsilon},$$

and $c < 1$ be such that:

$$(1 - c)E[B] = \Delta\epsilon/4.$$

Since $E[B] > \Delta\epsilon/2$, we have $c \geq 1/2$. Using Chernoff bound, we have that:

$$\begin{aligned}\mathbf{Pr}(B \leq \Delta\epsilon/4) &\leq e^{-\frac{c^2 E[B]}{2}} \\ &\leq e^{-\frac{1}{4}\frac{8\log(1/\delta)}{2}} \leq \delta.\end{aligned}$$

Thus $\mathbf{Pr}(b_1(\hat{s}_{\min}) + b_2(\hat{s}_{\min}) > \epsilon) \leq \delta$. □

For each dataset $\mathcal{D}$ of Table I and for itemset sizes $k = 2, 3, 4$, we applied Algorithm 1 setting $\Delta = 1,000$ and $\epsilon = 0.01$. The values of $\hat{s}_{\min}$ we obtained are reported in Table II (we added the prefix "Rand" to each dataset name, to de-

Algorithm 1 FindPoissonThreshold

Input: Dataset $\mathcal{D}$ of t transactions over n items, vector $\vec{f}$ of item frequencies, k, Δ, ε;
Output: Estimate $\hat{s}_{\min}$ of $s_{\min}$;

1: $\tilde{s} \leftarrow$ highest expected support of a k-itemset;
2: $s_{\max} \leftarrow 0$;
3: $W \leftarrow \emptyset$;
4: **for** $i \leftarrow 1$ to Δ **do**
5: $\quad \hat{\mathcal{D}}_i \leftarrow$ random dataset with parameters t,n,$\vec{f}$;
6: $\quad W \leftarrow W \cup \left\{\text{frequent } k\text{-itemsets in } \hat{\mathcal{D}}_i \text{ w.r.t. } \tilde{s}\right\}$;
7: **if** $W = \emptyset$ **then**
8: $\quad \tilde{s} \leftarrow \tilde{s}/2$;
9: $\quad$ **goto** 4;
10: **if** $(s_{\max} = 0)$ **then**
11: $\quad s_{\max} \leftarrow \max_{X \in W, \hat{\mathcal{D}}_i} \left\{\text{support of } X \text{ in } \hat{\mathcal{D}}_i\right\} + 1$;
12: **for** $s \leftarrow \tilde{s}$ to $s_{\max}$ **do**
13: $\quad$ **for all** $X \in W$ **do**
14: $\quad\quad p_X(s) \leftarrow$ empirical probability of $\{Z_{X,s} = 1\}$;
15: $\quad$ **for all** $X, Y \in W : X \cap Y \neq \emptyset$ **do**
16: $\quad\quad p_{X,Y}(s) \leftarrow$ empirical probability of $\{Z_{X,s} Z_{Y,s} = 1\}$;
17: $\quad b_1(s) \leftarrow \sum_{X,Y \in W; Y \in I(X)} p_X(s) p_Y(s)$;
18: $\quad b_2(s) \leftarrow \sum_{X,Y \in W; X \neq Y \in I(X)} p_{X,Y}(s)$;
19: **if** $b_1(\tilde{s}) + b_2(\tilde{s}) \leq \varepsilon/4$ **then**
20: $\quad s_{\max} \leftarrow \tilde{s}$;
21: $\quad \tilde{s} \leftarrow \tilde{s}/2$;
22: $\quad$ **goto** 3;
23: $\hat{s}_{\min} \leftarrow \min\{s > \tilde{s} : b_1(s) + b_2(s) \leq \varepsilon/4\}$;
24: **return** $\hat{s}_{\min}$;

note the fact that the dataset is random and features the same parameters as the corresponding real one).

3. PROCEDURES FOR THE DISCOVERY OF HIGH-SUPPORT SIGNIFICANT ITEMSETS

For a given itemset size k, the value $s_{\min}$ identifies a region of (relatively high) supports where we concentrate our quest for statistically significant k-itemsets. In this section we develop procedures to identify a family of k-itemsets (among those of support greater than or equal to $s_{\min}$) which are statistically significant with a controlled FDR. More specifically, in Subsection 3.1 we show that a family with the desired properties can be obtained as a subset of the frequent k-itemsets with

	$\hat{s}_{\min}$		
Dataset	$k=2$	$k=3$	$k=4$
RandRetail	9237	4366	784
RandKosarak	273266	100543	20120
RandBms1	268	23	5
RandBms2	168	13	4
RandBmspos	76672	15714	2717
RandPumsb*	29303	21893	16265

Table II. Values of $\hat{\boldsymbol{s}}_{\min}$ for $\epsilon = 0.01$ and for $k = 2, 3, 4$, in random datasets with the same values of n, t, and with the same frequencies of the items as the corresponding benchmark datasets.

respect to $s_{\min}$, selected based on a standard multi-comparison test. However, the size of the returned family may be rather small, due to the correction required to account for the possibly large number of null hypotheses, thus incurring a large number of false negatives. To achieve higher effectiveness, in Subsection 3.2 we devise a more sophisticated procedure which identifies a support threshold $s^* \geq s_{\min}$ such that *all* frequent k-itemsets with respect to s^* are statistically significant with a controlled FDR. In the next section we will provide experimental evidence that in many cases the latter procedure yields much fewer false negatives.

3.1 A procedure based on a standard multi-comparison test

We present a first, simple procedure to discover significant itemsets with controlled FDR, based on the following well established result in multi-comparison testing.

THEOREM 5 [BENJAMINI AND YEKUTIELI 2001]. *Assume that we are testing for m null hypotheses. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered observed p-values of the m tests. For a given parameter β, with $0 < \beta < 1$, define*

$$\ell = \max\left\{ i \geq 0 : p_{(i)} \leq \frac{i}{m \sum_{j=1}^{m} \frac{1}{j}} \beta \right\}, \tag{2}$$

and reject the null hypotheses corresponding to tests $(1), \ldots, (\ell)$. Then, the FDR for the set of rejected null hypotheses is upper bounded by β.

Let $\mathcal{D}$ denote an input dataset consisting of t transactions over n items, and let k be the fixed itemset size. Recall that $s_{\min}$ is the minimum support threshold for which the distribution of $\hat{Q}_{k,s}$ is well approximated by a Poisson distribution. First, we mine from $\mathcal{D}$ the set of frequent k-itemsets $\mathcal{F}_{(k)}(s_{\min})$. Then, for each $X \in \mathcal{F}_{(k)}(s_{\min})$, we test the null hypothesis H_0^X that the observed support of X in $\mathcal{D}$ is drawn from a Binomial distribution with parameters t and f_X (the product of the individual frequencies of the items of X), setting the rejection threshold as specified by condition (2), with parameters β and $m = \binom{n}{k}$. Based on Theorem 5, the itemsets of $\mathcal{F}_{(k)}(s_{\min})$ whose associated null hypothesis is rejected can be returned as significant, with FDR upper bounded by β. The pseudocode Procedure 1 implements the strategy described above. We want to remark that the application of Theorem 5 made in the procedure is conservative. In fact, while all itemsets of size k are considered by setting the number of null hypotheses $m = \binom{n}{k}$, p-values

are calculated only for itemsets in $\mathcal{F}_{(k)}(s_{\min})$. It is easy to argue that in this fashion the returned family of itemsets is a subset of the family the would be returned if p-values were calculated for all m itemsets, which, however, would be extremely time consuming even for small values of k, and would nullify the purpose of having a support threshold.

Procedure 1

Input: Dataset $\mathcal{D}$ of t transactions over n items, vector $\vec{f}$ of item frequencies, k, $\beta \in (0,1)$;
Output: Family of significant k-itemsets with FDR $\leq \beta$;
Determine $s_{\min}$ and compute $\mathcal{F}_{(k)}(s_{\min})$ from $\mathcal{D}$;
for all $X \in \mathcal{F}_{(k)}(s_{\min})$ **do**
$s_X \leftarrow$ support of X in $\mathcal{D}$;
$f_X \leftarrow \Pi_{i \in X} f_i$;
$p^{(X)} \leftarrow \mathbf{Pr}(\mathrm{Bin}(t, f_X) \geq s_X)$;
Let $p_{(1)}, p_{(2)}, \ldots,$ be the sorted sequence of the values $p^{(X)}$, with $X \in \mathcal{F}_{(k)}(s_{\min})$;
$m \leftarrow \binom{n}{k}$;
$\ell = \max\left\{0, i : p_{(i)} \leq \frac{i}{m \sum_{j=1}^{m} \frac{1}{j}} \beta\right\}$;
return $\left\{X \in \mathcal{F}_{(k)}(s_{\min}) : p^{(X)} = p_{(i)}, 1 \leq i \leq \ell\right\}$;

3.2 Establishing a support threshold for significant frequent itemsets

Let α and β be two constants in $(0,1)$. We seek a threshold s^* such that, with confidence $1-\alpha$, the k-itemsets in $\mathcal{F}_{(k)}(s^*)$ can be flagged as statistically significant with FDR at most β. The threshold s^* is determined through a robust statistical approach which ensures that the number $Q_{k,s^*} = |\mathcal{F}_{(k)}(s^*)|$ deviates significantly from what would be expected in a random dataset, and that the magnitude of the deviation is sufficient to guarantee the bound on the FDR.

Let $s_{\min}$ be the minimum support such that the Poisson approximation for the distribution of $\hat{Q}_{k,s}$ holds for $s \geq s_{\min}$, and let $s_{\max}$ be the maximum support of an item (hence, of an itemset) in $\mathcal{D}$. Our procedure performs $h = \lfloor \log_2(s_{\max} - s_{\min}) \rfloor + 1$ comparisons. Let $s_0 = s_{\min}$ and $s_i = s_{\min} + 2^i$, for $1 \leq i < h$. In the i-th comparison, with $0 \leq i < h$, we test the null hypothesis H_0^i that the observed value Q_{k,s_i} is drawn from the same Poisson distribution as $\hat{Q}_{k,s_i}$. We choose as s^* the minimum of the s_i's, if any, for which the null hypothesis H_0^i is rejected.

For the correctness of the above procedure, it is crucial to specify a suitable rejection condition for each H_0^i. Assume first that, for $0 \leq i < h$, we reject the null hypothesis H_0^i when the p-value of the observed value Q_{k,s_i} is smaller than α_i, where the α_i's are chosen so that $\sum_{i=0}^{h-1} \alpha_i = \alpha$. Then, the union bound shows that the probability of rejecting any true null hypothesis is less than α. However, this approach does not yield a bound on the FDR for the set $\mathcal{F}_{(k)}(s^*)$. In fact, some itemsets in $\mathcal{F}_{(k)}(s^*)$ are likely to occur with high support even under H_0^i, hence they would represent false discoveries. The impact of this phenomenon can be contained

by ensuring that the FDR is below a specified level β. To this purpose, we must strengthen the rejection condition, as explained below.

Fix suitable values $\beta_0, \beta_1, \ldots, \beta_{h-1}$ such that $\sum_{i=0}^{h-1} \beta_i^{-1} \leq \beta$. For $0 \leq i < h$, let $\lambda_i = E[\hat{Q}_{k,s_i}]$. We now reject H_0^i when the p-value of Q_{k,s_i} is smaller than α_i, *and* $Q_{k,s_i} \geq \beta_i \lambda_i$. The following theorem establishes the correctness of this approach.

THEOREM 6. *With confidence* $1 - \alpha$, $\mathcal{F}_{(k)}(s^*)$ *is a family of statistically significant frequent* k*-itemsets with FDR at most* β.

PROOF. Observe that since $\sum_{i=0}^{h-1} \alpha_i \leq \alpha$, we have that all rejections are correct, with probability at least $1-\alpha$. Let E_i be the event *"H_0^i is rejected"* or equivalently, *"the p-value of Q_{k,s_i} is smaller than α_i and $Q_{k,s_i} \geq \beta_i \lambda_i$"*. Suppose that H_0^i is the first rejected null hypothesis, for some index i, whence $s^* = s_i$ and Q_{k,s_i} itemsets are flagged as significant. Let V_i be the number of false discoveries among the Q_{k,s_i} itemsets of support at least s_i, given that H_0^i is the first rejected null hypothesis. The expectation of V_i is $E[X_i | E_i, \bar{E}_{i-1}, \ldots, \bar{E}_0]$, where X_i is a Poisson random variable with expectation λ_i. Since $Q_{k,s_i} \geq \beta_i \lambda_i$ when H_0^i is rejected, by the law of total probability we have

$$\begin{aligned}
FDR &\leq \sum_{i=0}^{h-1} E\left[\frac{V_i}{Q_{k,s_i}}\right] \mathbf{Pr}(E_i, \bar{E}_{i-1}, \ldots, \bar{E}_0) \\
&\leq \sum_{i=0}^{h-1} \frac{E[V_i]}{\beta_i \lambda_i} \mathbf{Pr}(E_i, \bar{E}_{i-1}, \ldots, \bar{E}_0) \\
&\leq \sum_{i=0}^{h-1} \frac{E[X_i \mid E_i, \bar{E}_{i-1}, \ldots, \bar{E}_0]}{\beta_i \lambda_i} \mathbf{Pr}(E_i, \bar{E}_{i-1}, \ldots, \bar{E}_0) \\
&= \sum_{i=0}^{h-1} \frac{\sum_{j \geq 0} j \mathbf{Pr}(X_i = j, E_i, \bar{E}_{i-1}, \ldots, \bar{E}_0)}{\beta_i \lambda_i} \\
&\leq \sum_{i=0}^{h-1} \frac{\lambda_i}{\beta_i \lambda_i} = \sum_{i=0}^{h-1} \frac{1}{\beta_i} \leq \beta.
\end{aligned}$$

□

The pseudocode Procedure 2 specifies more formally our approach to determine the support threshold s^*. Note that estimates for the λ_i's needed in the for-loop of Lines 7-9 can be obtained from the same random datasets generated in Algorithm 1, which are used there for the estimation of $s_{\min}$.

We remark that the number h of null hypotheses tested in our approach is kept small purposely, namely logarithmic in the maximum support of an itemset (hence in the number of transactions), so that only a moderate correction due to the multiplicity of hypotheses is required, at the expense of a reasonably controlled loss of precision in determining the support threshold s^*. In particular, the parameters α_i and β_i, with $0 \leq i < h$, can be set to values which are not excessively small, so that higher effectiveness can be achieved.

Procedure 2

Input: Dataset $\mathcal{D}$ of t transactions over n items, vector $\vec{f}$ of item frequencies, k, $\alpha, \beta \in (0,1)$;

Output: s^* such that, with confidence $1-\alpha$, $\mathcal{F}_{(k)}(s^*)$ is a family of significant k-itemsets with FDR $\leq \beta$;

1: Determine $s_{\min}$ and compute $\mathcal{F}_{(k)}(s_{\min})$ from $\mathcal{D}$;
2: $s_{\max} \leftarrow$ maximum support of an item;
3: $i \leftarrow 0$; $s_0 \leftarrow s_{\min}$;
4: $h \leftarrow \lfloor \log_2(s_{\max} - s_{\min}) \rfloor + 1$;
5: Fix $\alpha_0, \ldots, \alpha_{h-1} \in (0,1)$ s.t. $\sum_{i=0}^{h-1} \alpha_i = \alpha$;
6: Fix $\beta_0, \ldots, \beta_{h-1} \in (0,1)$ s.t. $\sum_{i=0}^{h-1} \beta_i^{-1} = \beta$;
7: **for** $i \leftarrow 0$ to $h-1$ **do**
8: Compute $\lambda_i = E[\hat{Q}_{k,s_i}]$;
9: **while** $i < h$ **do**
10: Compute Q_{k,s_i};
11: **if** ($\mathbf{Pr}(\text{Poisson}(\lambda_i) \geq Q_{k,s_i}) \leq \alpha_i$) **and** ($Q_{k,s_i} \geq \beta_i \lambda_i$) **then**
12: **return** $s^* \leftarrow s_i$;
13: $s_{i+1} \leftarrow s_{\min} + 2^{i+1}$;
14: $i \leftarrow i+1$;
15: **return** $s^* \leftarrow \infty$;

Example: To demonstrate the power of Procedure 2, and to compare it to the standard FDR method, Procedure 1, we consider a simple dataset of 10^6 transactions over 10^4 items and frequency 10^{-3} per item. The dataset has 600 disjoint pairs that each appear in exactly 10 transactions, The remaining 8800 items are placed randomly in the transactions. We compare the outcome of the two procedures when testing for statistically significant frequent itemsets of size 2, with FDR $\beta = 0.1$.

Consider first the standard FDR test, Procedure 1. The p-value for the hypothesis that a given pair appears in at least 10 transactions is $\binom{10^6}{10}(10^{-3})^{20} > 1.1 \times 10^{-7}$. Since there are $\binom{10^4}{2}$ hypothesis, Procedure 1 does not flag any pair of items as statistically significant.

Consider now the application of our method to this input. Using Algorithm 1 we derive $s_{\min} = 10$. Thus, Procedure 2 tests all pairs with support at least 10. The expected number of such pairs among the items placed at random is less than 6. Since the maximum support of an item is 10^3, 10 hypothesis are tested. By the Poisson the approximation, the probability of observing 606 pairs with support at least 10 in the null hypothesis is extremely small ($< 10^{-32}$), and thus Procedure 2 flags the 606 pairs with supports at least 10 with $\alpha = 0.05$, correctly reporting all the significant itemsets and, and the expected number of false positive is no more than 6, thus the FDR is indeed bounded by $\beta = 0.1$.

4. EXPERIMENTAL RESULTS

In order to show the potential of our approach, in this section we report on a number of experiments performed on the benchmark datasets of Table I. First, in Subsection 4.1, we validate experimentally the methodology implemented by Procedure 2, while in Subsection 4.2, we compare Procedure 2 against the more standard Procedure 1, with respect to their ability to discover significant itemsets.

4.1 Experiments on benchmark datasets

For each benchmark dataset in Table I and for $k = 2, 3, 4$, we apply Procedure 2 with $\alpha = \beta = 0.05$, and $\alpha_i = \beta_i^{-1} = 0.05/h$. The results are displayed in Table III, where, for each dataset and for each value of k, we show: the support s^* returned by Procedure 2, the number Q_{k,s^*} of k-itemsets with support at least s^*, and the expected number $\lambda(s^*)$ of itemsets with support at least s^* in a corresponding random dataset.

	$k=2$			$k=3$			$k=4$		
Dataset	s^*	Q_{k,s^*}	$\lambda(s^*)$	s^*	Q_{k,s^*}	$\lambda(s^*)$	s^*	Q_{k,s^*}	$\lambda(s^*)$
Retail	∞	0	0	∞	0	0	848	6	0.01
Kosarak	∞	0	0	∞	0	0	21144	12	0.01
Bms1	276	56	0.19	23	258859	0.06	5	27M	0.05
Bms2	168	429	0.73	13	36112	0.25	4	714045	0.01
Bmspos	∞	0	0	16226	22	0.01	2717	891	0.38
Pumsb*	29303	29	0.05	21893	406	0.35	16265	6293	1.37

Table III. Results obtained by applying Procedure 2 with $\alpha = 0.05, \beta = 0.05$ and $k = 2, 3, 4$ to the benchmark datasets of Table I.

We observe that for most pairs (dataset,k) the number of significant frequent k-itemsets obtained is rather small, but, in fact, at support s^* in random instances of those datasets, less than two (often much less than one) frequent k-itemsets would be expected. These results provide evidence that our methodology not only defines significance on statistically rigorous grounds, but also provides the mining task with suitable support thresholds that avoid explosion of the output size (the widely recognized "Achilles' heel" of traditional frequent itemset mining). This feature crucially relies on the identification of a region of "rare events" provided by the Poisson approximation. As discussed in Section 1.3, the discovery of significant itemsets with low support (not returned by our method) would require the extraction of a large (possibly exponential) number of itemsets, that would make any strategy aiming to discover these itemsets unfeasible. Instead, we provide an efficient method to identify, with high confidence level, the family of most frequent itemsets that are statistically significant without overwhelming the user with a huge number of discoveries.

There are, however, a few cases where the number of itemsets returned is still considerably high. Their large number may serve as a sign that the results call for further analysis, possibly using clustering techniques [Xin et al. 2005] or limiting

the search to *closed itemsets* [Pasquier et al. 1999][1]. For example, consider dataset Bms1 with $k = 4$ and the corresponding value $s^* = 5$ from Table III. Extracting the closed itemsets of support greater or equal to s^* in that dataset revealed the presence of a closed itemset of cardinality 154 with support greater than 7 in the dataset. This itemset, whose occurrence by itself represents an extremely unlikely event in a random dataset, accounts for more than 22M non-closed subsets with the same support among the 27M reported as significant.

It is interesting to observe that the results obtained for dataset Retail provide further evidence for the conclusions drawn in [Gionis et al. 2006], which suggested random behavior for this dataset (although the random model in that work is slightly different from ours, in that the family of random datasets also maintains the same transaction lengths as the real one). Indeed, no support threshold s^* could be established for mining significant k-itemsets with $k = 2, 3$, while the support threshold s^* identified for $k = 4$ yielded as few as 6 itemsets. However, the conclusion drawn in [Gionis et al. 2006] was based on a qualitative assessment of the discrepancy between the numbers of frequent itemsets in the random and real datasets, while our methodology confirms the findings on a statistically sound and rigorous basis.

Observe also that for some other pairs (dataset,k) our procedure does not identify any support threshold useful for mining statistically significant itemsets. This is an evidence that, for the specific k and for the high supports considered by our approach, these datasets do not present a significant deviation from the corresponding random datasets.

Finally, in order to assess the validity of our methodology we applied it to random datasets. Specifically, for each benchmark dataset of Table I and for $k = 2, 3, 4$, we generated 100 random instances with the same parameters as those of the benchmark, and applied Procedure 2 to each instance, searching for a support threshold s^* for mining significant itemsets. In Table IV we report the number of times Procedure 2 was successful in returning a finite value for s^*. As expected, the procedure returned $s^* = \infty$, in *all cases* but for 2 of the 100 instances of the random dataset with the same parameters as dataset Pumsb* with $k = 2$. However, in these two latter cases, mining at the identified support threshold only yielded a very small number of significant itemsets (one and two, respectively).

4.2 Relative effectiveness of Procedures 1 and 2

In order to assess the relative effectiveness of the two procedures presented in the previous section, we applied them to the benchmark datasets of Table I. Specifically, we compared the number of itemsets extracted using the threshold s^* provided by Procedure 2, with the number of itemsets flagged as significant using the more standard method based on Benjamini and Yekutieli's technique (Procedure 1), imposing the same upper bound $\beta = 0.05$ on the FDR.

The results are displayed in Table V, where for each pair (dataset,k), we report

[1] An itemset is *closed* if it is not properly contained in another itemset with the same support.

	$s^* < \infty$		
Dataset	$k = 2$	$k = 3$	$k = 4$
RandomRetail	0	0	0
RandomKosarak	0	0	0
RandomBms1	0	0	0
RandomBms2	0	0	0
RandomBmspos	0	0	0
RandomPumsb*	2	0	0

Table IV. Results for Procedure 2 with $\alpha = 0.05, \beta = 0.05$ for random versions of benchmark datasets; each entry reports the number of times, out of 100 trials, the procedure returned a finite value for s^*.

the cardinality of the family $\mathcal{R}$ of k-itemsets flagged as significant by Procedure 1, and the ratio $r = Q_{k,s^*}/|\mathcal{R}|$, where Q_{k,s^*} is the number of k-itemsets of support at least s^*, which are returned as significant with the methodology of Subsection 3.2.

We observe that in all cases where Procedure 2 returned a finite value of s^* the ratio r is greater than or equal to 1 (except for dataset Bms1 and $k = 2$, and dataset Bmspos and $k = 3$, where r is however very close to 1). Moreover, in some cases the ratio r is rather large. Since both methodologies identify significant k-itemsets among all those of support at least $s_{\min}$, these results provide evidence that the methodology of Subsection 3.2 is often more (sometimes much more) effective. The methodology succeeds in identifying more significant itemsets, since it evaluates the significance of the *entire* set $\mathcal{F}_{(k)}(s^*)$ by comparing Q_{k,s^*} to $\hat{Q}_{k,s^*}$. In contrast, Procedure 1 must implicitly test considerably more hypotheses (corresponding to the significance all possible k-itemsets), thus the power of the test (1-Pr(Type-II error)) is significantly smaller.

Observe that the cases where $r = 0$ in Table V correspond to pairs (dataset,k) for which Procedure 2 returned $s^* = \infty$, that is, the procedure was not able to identify a threshold for mining significant k-itemsets. Note, however, that in all of these cases the number of significant k-itemsets returned by Procedure 1 is extremely small (between 1 and 3). Hence, for these pairs, both methodologies indicate that there is very little significant information to be mined at high supports.

	$k = 2$		$k = 3$		$k = 4$	
Dataset	$\lvert\mathcal{R}\rvert$	r	$\lvert\mathcal{R}\rvert$	r	$\lvert\mathcal{R}\rvert$	r
Retail	3	0	3	0	6	1.0
Kosarak	1	0	1	0	12	1.0
Bms1	60	0.933	64367	4.441	219706	122.9
Bms2	429	1.0	25906	1.394	60927	11.72
Bmspos	2	0	23	0.957	891	1.0
Pumsb*	29	1.0	406	1.0	6288	1.001

Table V. Results using Test 1 to bound the FDR with $\beta = 0.05$ for itemsets of support $\geq s_{\min}$.

5. CONCLUSIONS

The main technical contribution of this work is the proof that in a random dataset where items are placed independently in transactions, there is a minimum support $s_{\min}$ such that the number of k-itemsets with support at least $s_{\min}$ is well approximated by a Poisson distribution. The expectation of the Poisson distribution and the threshold $s_{\min}$ are functions of the number of transactions, number of items, and frequencies of individual items.

This result is at the base of a novel methodology for mining frequent itemsets which can be flagged as statistically significant incurring a small FDR. In particular, we use the Poisson distribution as the distribution under the null hypothesis in a novel multi-hypothesis statistical approach for identifying a suitable support threshold $s^* \geq s_{\min}$ for the mining task. We control the FDR of the output in a way which takes into account global characteristics of the dataset, hence it turns out to be more powerful than other standard statistical tools (e.g., [Benjamini and Yekutieli 2001]). The results of a number of experiments, reported in the paper, provide evidence of the effectiveness of our approach.

To the best of our knowledge, our methodology represents the first attempt at establishing a support threshold for the classical frequent itemset mining problem with a quantitative guarantee on the significance of the output.

We note that our method is most powerful when all items have similar frequencies. When there are large differences between the frequencies of individual items there can be itemsets with small expected support in the corresponding random dataset. Such itemsets can have statistically significant support in the actual dataset although their supports are not above the threshold for the Poisson approximation, and they are not among the most frequent itemsets. Our current technique does not flag such discoveries. A possible solution for this problem that requires further research is to stratify the observed data according to frequencies and analyze the statistical significance in each group separately.

REFERENCES

Aggarwal, C. and Yu, P. 1998. A new framework for itemset generation. In *Proc. of the 17th ACM Symp. on Principles of Database Systems*. 18–24.

Agrawal, R., Imielinski, T., and Swami, A. 1993. Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD Intl. Conference on Management of Data*. 207–216.

Arratia, R., Goldstein, L., and Gordon, L. 1990. Poisson approximation and the Chen-Stein method. *Statistical Science 5,* 4, 403–434.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate. *Journal of the Royal Statistical Society,* Series B 57, 289–300.

Benjamini, Y. and Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics 29,* 4, 1165–1188.

Bolton, R., Hand, D., and Adams, N. 2002. Determining hit rate in pattern search. In *Proc. of Pattern Detection and Discovery*. LNAI 2447. 36–48.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. 2003. Multiple hypothesis testing in microarray experiments. *Statistical Science 18,* 1, 71–103.

DuMouchel, W. 1999. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The American Statistician 53*, 177–202.

DuMouchel, W. and Pregibon, D. 2001. Empirical bayes screening for multi-item associations. In *Proc. of the 7th*. 67–76.

Gionis, A., Mannila, H., Mielikäinen, T., and Tsaparas, P. 2006. Assessing data mining results via swap randomization. In *Proc. of the 12th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*. 167–176.

Goethals, B., Bayardo, R., and Zaki, M. J., Eds. 2004. *Proc. of the 2nd Workshop on Frequent Itemset Mining Implementations (FIMI04)*. Vol. 126. CEUR-WS Workshop On-line Proceedings.

Goethals, B. and Zaki, M. J., Eds. 2003. *Proc. of the 1st Workshop on Frequent Itemset Mining Implementations (FIMI03)*. Vol. 90. CEUR-WS Workshop On-line Proceedings.

Hämäläinen, W. and Nykänen, M. 2008. Efficient discovery of statistically significant association rules. In *Proc. of the 8th IEEE Intl. Conference on Data Mining*. 203–212.

Han, J., Cheng, H., Xin, D., and Yan, X. 2007. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery 15,* 1, 55–86.

Han, J. and Kamber, M. 2001. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Mateo, CA.

Jaroszewicz, S. and Scheffer, T. 2005. Fast discovery of unexpected patterns in data, relative to a bayesian. In *Proc. of the 11th Intl. Conference on Knowledge Discovery and Data Mining*. 118–127.

Megiddo, N. and Srikant, R. 1998. Discovering predictive association rules. In *Proc. of the 4thIntl. Conference on Knowledge Discovery and Data Mining*. 274–278.

Mitzenmacher, M. and Upfal, E. 2005. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, Cambridge MA.

Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. 1999. Discovering frequent closed itemsets for association rules. In *Proc. of the 7th Int. Conference on Database Theory*. 398–416.

Purdom, P. W., Gucht, D. V., and Groth, D. P. 2004. Average case performance of the apriori algorithm. *SIAM Journal on Computing 33,* 5, 1223–1260.

Sayrafi, B., Gucht, D. V., and Purdom, P. W. 2005. On the effectiveness and efficiency of computing bounds on the support of item-sets in the frequent item-sets mining problem. In *Proc. of the 1st International Workshop on Open Source Data Mining*. 46–55.

Silberschatz, A. and Tuzhilin, A. 1996. What makes patterns interesting in knowledge discovery systems. *IEEE Trans. on Knowledge and Data Engineering 8,* 6, 970–974.

Silverstein, C., Brin, S., and Motwani, R. 1998. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery 2,* 1, 39–68.

Srikant, R. and Agrawal, R. 1996. Mining quantitative association rules in large relational tables. In *Proc. of the ACM SIGMOD Intl. Conference on Management of Data*. 1–12.

Tan, P., Steinbach, M., and Kumar, V. 2006. *Introduction to Data Mining*. Addison Wesley.

Xin, D., Han, J., Yan, X., and Cheng, H. 2005. Mining compressed frequent-pattern sets. In *Proc. of the 31st Very Large Data Base Conference*. 709–720.

Zhang, H., Padmanabhan, B., and Tuzhilin, A. 2004. On the discovery of significant statistical quantitative rules. In *Proc. of the 10th Intl. Conference on Knowledge Discovery and Data Mining*. 374–383.