An Efficient Rigorous Approach for Identifying Statistically Significant Frequent Itemsets

Adam Kirsch[*]
Harvard University

Michael Mitzenmacher[†]
Harvard University

Andrea Pietracaprina[‡]
Univesity of Padova

Geppino Pucci[§]
University of Padova

Eli Upfal[¶]
Brown University

Fabio Vandin[‖]
University of Padova

ABSTRACT

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, for which the number of itemsets with support at least s^* yields 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.

Our methodology hinges on a Poisson approximation; we show that the distribution of the number of itemsets with support greater than some appropriate threshold $s_{\min}$ is approximately Poisson in a random dataset. We obtain this result through a novel application of the Chen-Stein approximation method, which is of independent interest. Based on this approximation, we develop an efficient parametric multi-hypothesis test for identifying the threshold s^*. A crucial feature of this approach is that, unlike most previous work, it takes into account the entire dataset rather than individual discoveries. It is therefore able to better distinguish between significant observations and random fluctuations. We present extensive experimental results to substantiate the effectiveness of our methodology.

1. Introduction

The discovery of frequent itemsets in transactional datasets is regarded as a fundamental primitive that arises in the mining of association rules and in many other scenarios [16,24]. In its original formulation, the problem requires that given a dataset $\mathcal{D}$ of transactions over a set of items $\mathcal{I}$, and a support threshold s, all itemsets $X \subseteq \mathcal{I}$ with support at least s (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. [2], a vast literature has flourished proposing variants of the problem, studying foundational issues, and presenting novel algorithmic strategies or clever implementations of known strategies (see, e.g., [12,13]), but many problems remain open [15]. 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 [16,24], 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 statistically significant frequent itemsets. 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 with no correlations between individual items. We also enforce that the returned family of significant itemsets feature

[*] Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA. Email: kirsch@eecs.harvard.edu

[†] Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA. Email: michaelm@eecs.harvard.edu

[‡] Department of Information Engineering, University of Padova, Italy. Email: andrea.pietracaprina@unipd.it

[§] Department of Information Engineering, University of Padova, Italy. Email: geppino.pucci@unipd.it

[¶] Computer Science Department, Brown University, Providence, RI, USA. Email: eli@cs.brown.edu

[‖] Department of Information Engineering, University of Padova, Italy. Email: vandinfa@dei.unipd.it

a small False Discovery Rate (FDR) [4].

1.1 The model

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 independently in transactions. 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 $f_i = n(i)/n$ 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. Among all possible $\binom{n}{k}$ itemsets of size k (k-*itemsets*) we are interested in statistically significant ones, that is, itemsets whose supports are significantly higher, in a statistical sense, than their expected supports in a dataset where individual items are placed independently in transactions.

Following [22], we consider a probability space of datasets with 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, independent of all other items and all other transactions.

Let $\hat{\mathcal{D}}$ denote a random dataset from this probability space. Given a support of an itemset in $\mathcal{D}$, the null hypothesis H_0 is that this support is drawn from the distribution of $\hat{D}$. The alternative hypothesis H_1 is that the support is not drawn from that distribution, and in particular that there is a positive correlation between the occurrences of the individual items in that itemset. An alternative H_0 model, proposed in [11], considers a sample space of all arrangements of n items to m transactions that satisfies the exact item frequencies and transaction lengths as $\mathcal{D}$. Conceivably, the technique of this paper could be adapted to this latter model as well.

1.2 Multi-hypothesis testing

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 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 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 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 pairs each with support at least 7. The probability of that event in the random dataset is less than 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.

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 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. In the context of frequent itemsets, if we test for significant itemsets of size k, then we are testing simultaneously $\binom{n}{k}$ null hypotheses. Each null hypothesis corresponds to the support of a given itemset not being statistically significant. 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 [8] for a good review).

The *False Discovery Rate (FDR)* was suggested by Benjamini and Hochberg [4] 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 $FDR = E[V/R]$ to be the expected ratio of erroneous rejections among all rejections (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 [4], Benjamini and Yekutieli [5] developed a general technique for controlling the FDR in any multi-hypothesis test (see Theorem 3).

1.3 Our Results

In this paper 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 can guarantee, in a statistical sense, that the returned family of frequent itemsets contains significant ones with a limited false discovery rate. 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 a corresponding random dataset according to the our random model described in Section 1.1. Let $Q_{k,s}$ be the number of itemsets of size k with support at least s with respect to $\mathcal{D}$, and let $\hat{Q}_{k,s}$ be the corresponding random variable for $\hat{\mathcal{D}}$. We show that there exists a minimum support value $s_{\min}$ (that depends on the parameters of $\mathcal{D}$), 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 [3].

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 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 confidence level α, the number of itemsets with support at least s^* is not sampled from a Poisson distribution and is therefore statistically significant. The fact that the number of itemsets with support at least s^* is statistically significant 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 expected ratio of false discoveries among all discoveries (FDR).

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 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 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 [5], and provide evidence that the latter often returns fewer significant itemsets, which indicates that our method has 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 [15, Section 3] for further references. The paper [1] relates 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 [1], is more effective than absolute support. In [9,10,23], 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, [9,10] proposes smoothing the ratio R using an empirical Bayesian approach. Bayesian analysis is also employed in [21] to derive subjective measures of significance of patterns (e.g., itemsets) based on how strongly they "shake" a system of established beliefs.

A statistical approach for identifying significant itemsets is presented in [22], 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 [9,10], there are technical flaws in the applications of the statistical test in [22]. Nevertheless, [22] 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 [11] propose a Markov chain-based approach 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 establish whether the real datasets exhibit 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 when mining signif-

Dataset	n	$[f_{\min}; f_{\max}]$	m	t
Retail	16470	[1.13e-05 ; 0.57]	10.3	88162
Kosarac	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 1: Parameters of the benchmark datasets: n is the number of items; $[f_{\min}, f_{\max}]$ is the range of frequencies of the individual items; m is the average transaction length; and t is the number of transactions.

icant patterns is studied in [6]. 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 [14], 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 [18] provide the means of evaluating the FDR in $\mathcal{A}$. The FDR metric is also employed in [26] in the context of discovering 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 1. 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 the Poisson approximation result for the random variable $\hat{Q}_{k,s}$. The statistical tests are presented in Section 3, and experimental results are reported in Section 4. Section 5 ends the paper with some concluding remarks.

2. Poisson Approximation for $\hat{Q}_{k,s}$

The Chen-Stein method [3] 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 Bernoulli random variables $\{Z_X \mid X \subset \mathcal{I},\ |X| = k\}$, such that $Z_X = 1$ if the itemset X appears in at least s transactions in the random dataset $\hat{\mathcal{D}}$, and $Z_X = 0$ otherwise. Also, let $p_X = \mathbf{Pr}(Z_X = 1)$. We are interested in the distribution of $\hat{Q}_{k,s} = \sum_{X:|X|=k} Z_X$.

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 and Z_X are independent. Adapting [3, Theorem 1] to our case we have:

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*

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

where

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

and

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

It is easy to see that the quantities b_1 and b_2 in the above theorem are both decreasing in s. Therefore, if $b_1 + b_2 < \epsilon$ for a given s, then the same upper bound will hold for every $s' > s$. Consequently, a given Poisson approximation for $\hat{Q}_{k,s}$, established through the above theorem, extends to $\hat{Q}_{k,s'}$ with $s' > s$.

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.

Theorem 2. *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. If*

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

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

$$\begin{aligned}\left\|\mathcal{L}(\hat{Q}_{k,s}) - \mathcal{L}(U)\right\| &= \sup_A |\mathbf{Pr}(\hat{Q}_{k,s} \in A) - \mathbf{Pr}(U \in A)| \\ &= O(1/n).\end{aligned}$$

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

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

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

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

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

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

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, let $g = |X \cap Y|$, and suppose that $g > 0$. Then if $Z_X Z_Y = 1$, 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,

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

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

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

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 $a \geq 2$)

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

Thus, $b_2 = O(n^{2k+sc-ak+(k-1)\left(\frac{a}{2}-1\right)})$. (Here we are using the fact that our choice of c satisfies $c \leq (k-1)(a-2)/2s$ to ensure that $n^{(k-1)\left(\frac{a}{2}-1\right)-cs} = \Omega(1)$.)

Now, we have $b_1 = O(1/n)$ since $c \leq (k-1)(a-2)/2s \leq k(a-2)/2s$, and $b_2 = O(1/n)$ since $c \leq [k(a-2) + (a-4)]/2s$. Thus, $b_1 + b_2 = O(1/n)$. □

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

The above section gives a rigorous analytical proof that there exists a meaningful range for the support s such that the number of itemsets of size k with support s or larger can be approximated by a Poisson variable. In practice, in order to avoid the inevitable slack due to the use of asymptotics in Theorem 2, we establish the minimum support $s_{\min}$ for the validity of the Poisson approximation via a simple Monte Carlo simulation which estimates the values b_1 and b_2 as defined in Theorem 1.

For clarity, we use the notation $b_1(s)$ and $b_2(s)$ to indicate explicitly that both quantities are functions of the support s. Suppose that for a chosen ϵ, $0 < \epsilon < 1$, we want to determine $s_{\min} = \min\{s \geq 1 \ : \ b_1(s) + b_2(s) \leq \epsilon\}$. Let $\tilde{s}$ be the maximum expected support of any k-itemset (we expect $\tilde{s} < s_{\min}$). We generate Δ random datasets and extract, from each such dataset, all k-itemsets of support at least $\tilde{s}$. Let W be the set of itemsets extracted in this fashion from all the generated datasets. It is easy to see that 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 $Z_X = 1$, and for each pair $X, Y \in W$, with $X \cap Y \neq \emptyset$, the empirical probability $p_{X,Y}$ of the event $(Z_X = 1) \wedge (Z_Y = 1)$. Note that for itemsets not in W these probabilities are estimated as 0. Then, if it turns out that $b_1(\tilde{s}) + b_2(\tilde{s}) > \epsilon$, we let $s_{\min}$ be the minimum $s > \tilde{s}$ such that $b_1(s) + b_2(s) \leq \epsilon$. Otherwise, if $b_1(\tilde{s}) + b_2(\tilde{s}) \leq \epsilon$, we repeat the above procedure starting from $\tilde{s}/2$. It can be shown (details omitted for lack of space) that for $\Delta = O(\log(1/\delta)/\epsilon)$, the output $\tilde{s}$ of the Monte-Carlo process satisfies

$$Pr(b_1(\tilde{s}) + b_2(\tilde{s}) \leq \epsilon) \geq 1 - \delta.$$

We remark that the information mined from the random datasets can also be used to obtain an estimate $\hat{\lambda}(s)$ of the expected number $\lambda(s)$ of itemsets with support at least s for every $s \geq s_{\min}$, that is the expectation of the Poisson distribution of $\hat{Q}_{k,s}$. We denote the set of values $\{\lambda(s) \ : \ s \geq s_{\min}\}$ by the vector $\overrightarrow{\lambda}$. These values are needed to perform our statistical tests illustrated in the next section.

For each dataset $\mathcal{D}$ of Table 1 and for itemset sizes $k = 2, 3, 4$, we determined the minimum value $s_{\min}$ so that the sum $b_1 + b_2$ is at most $\epsilon = 0.01$ for a corresponding random dataset $\hat{\mathcal{D}}$. The values of $s_{\min}$ we obtained are reported in Table 2 (we added the prefix "Rand" to each dataset name, to denote the fact that the dataset is random and features the same parameters as the corresponding real one).

3. Establishing a support threshold for mining statistically significant frequent itemsets

In this section we describe our testing methodology for determining a support threshold s^* such that the family of frequent itemsets with respect to s^* are statistically significant with a controlled FDR. At the end of the section (Subsection 3.1), we briefly describe a standard multi-comparison test to identify significant itemsets with small FDR, which will be employed in Section 4 as a comparison point to assess the benefits of our methodology.

Algorithm 1 FindPoissonThreshold

```
Input: $k$,$n$,$m$,$t$, $\Delta$, frequencies of individual items, $\varepsilon$.
Output: $s_{\min} = \min\{s \geq 1 \ : \ b_1(s) + b_2(s) \leq \epsilon\}$
 1: $\tilde{s} \leftarrow$ highest expected support of a $k$-itemset;
 2: $W \leftarrow \emptyset$;
 3: for $i \leftarrow 1$ to $\Delta$ do
 4:     $\hat{\mathcal{D}}_i \leftarrow$ random dataset with parameters $n$,$m$,$t$, and
        frequencies of individual items;
 5:     $W \leftarrow W \cup \left\{\text{frequent } k\text{-itemsets in } \hat{\mathcal{D}}_i \text{ w.r.t. } \tilde{s}\right\}$;
 6: if ($\tilde{s}$ == highest expected support of a $k$-itemset) then
 7:     $s_{\max} \leftarrow \max_{X \in W, \hat{\mathcal{D}}_i} \left\{\text{support of } X \text{ in } \hat{\mathcal{D}}_i\right\} + 1$;
 8: for all $s : \tilde{s} \leq s \leq s_{\max}$ do
 9:     for all $X \in W$ do
10:         $p_X(s) \leftarrow$ empirical probability of $\{Z_X = 1\}$;
11:     for all $X, Y \in W : X \cap Y \neq \emptyset$ do
12:         $p_{X,Y}(s) \leftarrow$ empirical probability of $\{Z_{X,Y} = 1\}$;
13:     $b_1(s) \leftarrow \sum_{X,Y \in W; Y \in I(X)} p_X(s) p_Y(s)$;
14:     $b_2(s) \leftarrow \sum_{X,Y \in W; X \neq Y \in I(X)} p_{X,Y}(s)$;
15: if $b_1(\tilde{s}) + b_2(\tilde{s}) \leq \varepsilon$ then
16:     $s_{\max} \leftarrow \tilde{s} - 1$;$\tilde{s} \leftarrow \tilde{s}/2$; goto 2;
17: $s_{\min} \leftarrow \min\{s > \tilde{s} : b_1(s) + b_2(s) \leq \varepsilon\}$;
18: return $s_{\min}$;
```

Let $\mathcal{D}$ be the input dataset and k a fixed itemset size. As before, we use $Q_{k,s}$ to denote the number of itemsets of size k of support at least s in $\mathcal{D}$, and $\hat{Q}_{k,s}$ to denote the corresponding random variable for $\hat{\mathcal{D}}$. We seek a support threshold s^* such that the observed value Q_{k,s^*} is significantly large compared to the expected value of $\hat{Q}_{k,s^*}$, in the sense that the probability that such a large deviation occurs in a random dataset is bounded by a pre-selected constant α.

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 in $\mathcal{D}$. Our testing methodology performs $h = \lfloor \log_2(s_{\max} - s_{\min}) \rfloor + 1$ comparisons. The null hypothesis H_0^i in the i-th comparison, for $0 \leq i < h$, is that the observed value Q_{k,s_i}, with $s_0 = s_{\min}$ and $s_i = s_{\min} + 2^i$ for $1 \leq i < h$, is drawn from the same Poisson distribution as $\hat{Q}_{k,s_i}$. For every $0 \leq i < h$ we fix a confidence level α_i and reject the null hypothesis H_0^i if the p-value of Q_{k,s_i} is smaller than α_i. If the α_i's are chosen so that $\sum_{i=0}^{h-1} \alpha_i = \alpha$, the union bound shows that the probability of rejecting any true null hypothesis is less than α.

Suppose we set s^* as the minimum of the s_i's for which the null hypothesis H_0^i was rejected. The probability that this number of itemsets of size k with support at least s_i were observed in a random dataset is bounded by α, thus

Dataset	$s_{\min}$		
	$k=2$	$k=3$	$k=4$
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 2: Minimum support $s_{\min}$ for which the Poisson approximation for $\hat{Q}_{k,s}$ holds (i.e., $b_1 + b_2 < 0.01$) for $k = 2, 3, 4$ in random datasets with the same values of n, t, m and with the same frequencies of the items as the corresponding benchmark datasets.

the size of this set is statistically significant. While this approach is a useful starting point in itself, it does not imply necessarily that all of these frequent itemsets are statistically significant. In fact, some of them are likely to occur with high support even under H_0^i, and hence they would represent false discoveries. In order to ensure that the FDR is below a specified level β, we can further amend the above procedure as follows.

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 modify the test described above by rejecting the i-th null hypothesis H_0^i if the probability that a Poisson random variable with expectation λ_i takes a value as large as the observed Q_{k,s_i} is at most α_i, *and* $Q_{k,s_i} \geq \beta_i \lambda_i$. Again, we set s^* as the minimum of the s_i's for which the null hypothesis H_0^i was rejected. We now prove that with this variation the FDR of the family of frequent itemsets of size k mined with threshold s^* is upper bounded by β.

We denote by $\mathcal{F}_{(k)}(s_i)$ the family of itemsets of size k with support at least s_i in $\mathcal{D}$, and note that $|\mathcal{F}_{(k)}(s_i)| = Q_{k,s_i}$. Let V_i be the number of false discoveries if H_0^i were the first null hypothesis rejected, in which case $\mathcal{F}_{(k)}(s_i)$ would be returned as the family of significant itemsets. Let E_i be the condition *"H_0^i is rejected"* or equivalently, *"the p-value of Q_{k,s_i} is smaller than α_i and $Q_{k,s_i} \geq \beta_i \lambda_i$"*. If a discovery is false positive then its distribution is as in the null hypotheses H_0^i. Thus, the distribution of V_i is stochastically bounded by that of a Poisson variable X_i with expectation λ_i, conditioned on the events $E_i, \bar{E}_{i-1}, \ldots, \bar{E}_0$. Therefore,

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

Since $\sum_{i=1}^{h} \alpha_i \leq \alpha$, and $\sum_{i=1}^{h} \beta_i^{-1} \leq \beta$ we obtain a test that identifies with confidence $1 - \alpha$ a minimum support s^* such that $|\mathcal{F}_{(k)}(s^*)|$ is significant. Moreover, the FDR among the individual itemsets of $\mathcal{F}_{(k)}(s^*)$ is bounded by β. The pseudocode Test 1 specifies more formally our test to determine the support threshold s^*.

Test 1

```
Input: $s_{\min}$, maximum item support $s_{\max}$, $\overrightarrow{\lambda}$,
    $\alpha_0, \ldots, \alpha_{h-1}$, with $\sum_{i=0}^{h-1} \alpha_i = \alpha$, and $\beta_0, \ldots, \beta_{h-1}$,
    with $\sum_{i=0}^{h-1} \beta_i^{-1} = \beta$
Output: $s^*$ such that $Q_{k,s^*}$ is significant with confidence
    $1 - \alpha$, and the FDR of the itemsets of $\mathcal{F}_{(k)}(s^*)$ is $\leq \beta$
Compute $\mathcal{F}_{(k)}(s_{\min})$;
$i \leftarrow 0$; $s_i \leftarrow s_{\min}$; $h \leftarrow \lfloor \log_2(s_{\max} - s_{\min}) \rfloor + 1$;
while $i < h$ do
    Compute $Q_{k,s_i}$;
    $p_{s_i} \leftarrow \mathbf{Pr}(\text{Poisson}(\lambda_i) \geq Q_{k,s_i})$;
    if $(p_{s_i} \leq \alpha_i)$ and $Q_{k,s_i} \geq \beta_i \lambda_i$ then
        return $s^* \leftarrow s_i$;
    $s_{i+1} \leftarrow s_{\min} + 2^{i+1}$; $i \leftarrow i + 1$;
return $s^* \leftarrow \infty$;
```

3.1 A standard multi-comparison test

To assess the effectiveness of our new approach we will compare it against the following standard multi-comparison test based on the state-of-the-art technique of [5].

THEOREM 3 (BENJAMINI AND YEKUTIELI [5]). *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 control of FDR at level β, define*

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

and reject the null hypotheses of tests $(1), \ldots, (\ell)$.

As before, let $\mathcal{D}$ denote the input dataset consisting of t transactions over n items. Let s be a given support threshold and k a fixed itemset size. After mining the frequent k-itemsets $\mathcal{F}_{(k)}(s)$ we test, for each $X \in \mathcal{F}_{(k)}(s)$, 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 (1), with parameters β and $m = \binom{n}{k}$. The itemsets of $\mathcal{F}_{(k)}(s)$ whose associated null hypothesis is rejected can be returned as significant with FDR upper bounded by β.

The pseudocode Test 2 summarizes the test described above.

4. Experimental Results

In this section, we report on a number of experiments devised to validate and show the potential of our approach.

Dataset	$k = 2$				$k = 3$				$k = 4$			
	s^*	Q_{k,s^*}	$\lambda(s^*)$	r	s^*	Q_{k,s^*}	$\lambda(s^*)$	r	s^*	Q_{k,s^*}	$\lambda(s^*)$	r
Retail	-	-	-	-	-	-	-	-	848	6	0.01	0.0017
Kosarak	-	-	-	-	-	-	-	-	21144	12	0.01	0.0008
Bms1	276	56	0.19	0.0034	23	258859	0.06	2×10^{-7}	5	27M	0.05	2×10^{-9}
Bms2	168	429	0.73	0.0017	13	36112	0.25	7×10^{-6}	4	714045	0.01	1×10^{-8}
Bmspos	-	-	-	-	16226	22	0.01	0.0005	2717	891	0.38	0.0004
Pumsb*	29303	29	0.05	0.0017	21893	406	0.35	0.0009	16265	6293	1.37	0.0002

Table 3: Results for Test 1 with $\alpha = 0.05, \beta = 0.05$.

```
Test 2
Input: $s$, $t$, vector $\vec{f}$ of frequencies of items in $\mathcal{D}$, $\beta$.
Output: family of significant itemsets with FDR $\leq \beta$;
  Compute $\mathcal{F}_{(k)}(s)$;
  $m \leftarrow \binom{n}{k}$;
  for all $X \in \mathcal{F}_{(k)}(s)$ do
    $s_X \leftarrow$ support of $X$ in $\mathcal{D}$;
    $f_X \leftarrow \Pi_{i \in X} f_i$;
    $p^{(X)} \leftarrow \mathbf{Pr}(\text{Bin}(t, f_X) \geq s_X)$;
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{p^{(X)}\}$;
  Let $p_{(1)}, p_{(2)}, \ldots,$ be the sorted sequence of the values
  $p^{(X)}$, with $X \in \mathcal{F}_{(k)}(s)$;
  $\ell = \max\left\{0, i : p_{(i)} \leq \frac{i}{m \sum_{j=1}^{m} \frac{1}{j}} \beta\right\}$;
  return $\left\{X : p^{(X)} = p_{(i)}, 1 \leq i \leq \ell\right\}$;
```

In Subsection 4.1, we apply our methodology to the benchmark datasets of Table 1. In Subsection 4.2, we compare the obtained results against those returned by the standard procedure to bound the FDR described in Subsection 3.1.

4.1 Experiments on benchmark datasets

We first apply our methodology to the benchmark datasets of Table 1. Specifically, for each dataset and for $k = 2, 3, 4$, we apply Test 1 to identify a support threshold s^* such that the number of k-itemsets that appear in at least s^* transactions represent a significant deviation from what would be expected in a random dataset, with significance level $\alpha = 0.05$, and with FDR of the returned family of itemsets at most $\beta = 0.05$. The results are displayed in Table 3, where, for each benchmark dataset and for $k = 2, 3, 4$, we show: the minimum support s^*, if any, for which the corresponding null hypothesis was rejected; the number of itemsets Q_{k,s^*} with support at least s^*; the expected number $\lambda(s^*)$ of itemsets with support at least s^* in a corresponding random dataset; and the ratio $r = \lambda(s^*)/Q_{k,s^*}$.

We also conducted an almost identical set of experiments, which are not reported in tabular form here for the sake of brevity, where we maintained the confidence level $\alpha = 0.05$ for the selection of s^*, but did not control the FDR (setting all β_i's to 0 in Test 1). These experiments behaved as one would expect; we observed some decrease in the support threshold required for rejection of the null hypothesis, and a consequent increase in the number of flagged itemsets, at the expense of some increase in the rate of false positive discoveries.

We observe that for most pairs (dataset,k) the number of significant frequent k-itemsets obtained is small, focusing on the most frequent significant itemsets. The 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 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 often serves as a sign that the results call for further analysis, possibly using clustering techniques [25] or searching for closed representatives for these itemsets [20]. For example, considering dataset Bms1 with $k = 4$ and the corresponding value $s^* = 5$ from Table 3. Extracting the closed itemsets of support greater or equal to s^* in that dataset revealed the presence of a closed itemset of cardinality 154 that appears more than 7 times in the dataset. This itemset, whose occurrence by itself represents an extremely unlikely event in a random dataset, accounts for more than 22M itemsets 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 [11], which suggests 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 only 6 itemsets. However, the conclusion drawn in [11] 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 tests do not find any support threshold useful to identify statistically significant itemsets. This is an evidence that, for the specific k and for the supports considered by our tests, these datasets do not present a significant deviation from a corresponding random dataset.

In order to assess its robustness, we applied our methodology to random datasets. Specifically, for each benchmark dataset of Table 1 and for $k = 2, 3, 4$, we generated 100 random instances with the same parameters as those of the benchmark, and applied Test 1 to each instance searching for a support threshold s^* for mining significant itemsets. As expected, the test was not able to reject any null hypothesis, hence it did not determine s^*, 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 cases, mining at the identified support threshold only yielded as few as 1 and 2 itemsets, respectively.

4.2 Comparison with the Standard FDR Test (Test 2)

We compare the number of itemsets extracted using the threshold s^* provided by Test 1, with the number of itemsets flagged as significant using the standard method described in Section 3.1 using the same threshold s^*. In both cases we imposed a bound $\beta = 0.05$ on the FDR.

The results are displayed in Table 4, where for each pair (dataset,k), we report the cardinality of the family $\mathcal{R}$ of significant itemsets returned by the method of Section 3.1, and the ratio $r = |\mathcal{R}|/Q_{k,s^*}$, where Q_{k,s^*} is the number of significant itemsets returned by our methodology. We observe that in some cases the method of Section 3.1 identifies a small fraction of the itemsets flagged as significant by Test 1. Indeed, in two cases more than 90% of the significant itemsets returned by our methodology are not flagged as significant by the standard method. Since we imposed a bound $\beta = 0.05$ on the FDR for both tests, the itemsets not identified by the method of Section 3.1 correspond in large part to significant itemsets. Test 1 succeeds in identifying these itemsets, since it evaluates the significance of the entire set of itemsets of support s^* comparing Q_{k,s^*} to $\hat{Q}_{k,s^*}$. In contrast, Test 2 has to control for testing of significantly more hypothesis (corresponding to the significance all possible k-itemsets), thus the power of the test (1-Pr(Type-II error)) is significantly smaller. These results demonstrate the advantage of our technique compared to the standard approach for multi-hypothesis tests that control FDR.

5. Conclusions

The main technical contribution of the paper 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.

Dataset	$k = 2$		$k = 3$		$k = 4$	
	$\|\mathcal{R}\|$	r	$\|\mathcal{R}\|$	r	$\|\mathcal{R}\|$	r
Retail	-	-	3	1.0	6	1.0
Kosarak	-	-	-	-	12	1.0
Bms1	60	0.984	64367	0.249	219706	0.008
Bms2	429	1.0	25906	0.717	60927	0.085
Bmspos	-	-	24	1.0	891	1.0
Pumsb*	29	1.0	406	1.0	6288	0.999

Table 4: Results using Test 2 to bound the FDR with $\beta = 0.05$ for itemsets of support $\geq s^*$.

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 of 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., [5]). 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.

6. References

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

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

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

[4] Y. Benjamini, and Y. Hochberg. Controlling the false discovery rate. *J. Royal Statistical Society*, Series B, 57:289–300, 1995.

[5] Y. Benjamini, and D. Yekutieli The control of the false discovery rate in multiple testing under dependency *Annals of Statistics*, 29 (4): 1165-1188, 2001

[6] R.J. Bolton, D.J. Hand, and N.M. Adams. Determining Hit Rate in Pattern Search In *Proc. of Pattern Detection and Discovery*, LNAI 2447, pages 36–48, 2002.

[7] W. J. Conover. *Practical Nonparametric Statistics*. Wiley Series in Probability, 3rd Ed., 1999.

[8] S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, Vol. 18, No. 1, 2003, p. 71-103.

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

[10] W. DuMouchel and D. Pregibon. Empirical Bayes screening for multi-item associations. In *Proc. of the 7th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, pages 67–76, 2001.

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

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

[13] B. Goethals and M. J. Zaki, editors. *Proc. of the 1st Workshop on Frequent Itemset Mining Implementations (FIMI03)*, volume 90. CEUR-WS Workshop On-line Proceedings, November 2003.

[14] W. Hämäläinen, and M. Nykänen Efficient discovery of statistically significant association rules In *Proc. of the 8th IEEE Intl. Conference on Data Mining*, 2008. To appear.

[15] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery*, 14(1), 2007.

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

[17] H.O. Lancaster. *The Chi-squared Distribution*. John Wiley & Sons, New York NY, 1969.

[18] N. Megiddo, and R. Srikant Discovering predictive association rules. In *Proc. of the 4th Intl. Conference on Knowledge Discovery and Data Mining*, pages 274–278, 1998.

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

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

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

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

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

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

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

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