Significance and Recovery of Block Structures in Binary Matrices with Noise

Xing Sun[1] and Andrew Nobel[1,2]

[1] Department of Statistics and Operation Research
[2] Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA
{xingsun, nobel}@email.unc.edu

Abstract. Frequent itemset mining (FIM) is one of the core problems in the field of Data Mining and occupies a central place in its literature. One equivalent form of FIM can be stated as follows: given a rectangular data matrix with binary entries, find every submatrix of 1s having a minimum number of columns. This paper presents a theoretical analysis of several statistical questions related to this problem when noise is present. We begin by establishing several results concerning the extremal behavior of submatrices of ones in a binary matrix with random entries. These results provide simple significance bounds for the output of FIM algorithms. We then consider the noise sensitivity of FIM algorithms under a simple binary additive noise model, and show that, even at small noise levels, large blocks of 1s leave behind fragments of only logarithmic size. Thus such blocks cannot be directly recovered by FIM algorithms, which search for submatrices of all 1s. On the positive side, we show how, in the presence of noise, an error-tolerant criterion can recover a square submatrix of 1s against a background of 0s, even when the size of the target submatrix is very small.

1 Introduction

Frequent itemset mining (FIM) [1, 2], also known as market basket analysis, is a central and well-studied problem in the field of Data Mining, and occupies a central place in its literature. It is closely related to a variety of related, more general problems, such as bi-clustering and subspace clustering [28, 7, 3, 10] that are of active interest to the Data Mining community. A variety of applications using FIM and other related bi-clustering algorithms can be found in [21, 14]. In the FIM problem the available data is described by a list $S = \{s_1, \ldots, s_n\}$ of items and a set $T = \{t_1, \ldots, t_m\}$ of transactions. Each transaction t_i consists of a subset of the items in S. (If S contains the items available for purchase at a store, then each t_i represents a record of items purchased during one transaction, without multiplicity.) The goal of FIM is to identify sets of items that appear together in more than k transactions, where $k \geq 1$ is a threshold for "frequent". The data for the FIM problem can readily be represented by an $m \times n$ binary

matrix $\mathbf{X}$, with $x_{i,j} = 1$ if transaction t_i contains item s_j, and $x_{i,j} = 0$ otherwise. In this form the FIM problem can be stated as follows: given $\mathbf{X}$ and $k \geq 1$, find every submatrix of 1s in $\mathbf{X}$ having at least k columns. Frequent itemset algorithms perform an exhaustive search for such submatrices.

The application of FIM to large data sets for the purposes of exploratory analysis raises a number of natural statistical questions. In this paper we present (preliminary) answers to three such questions. The first question considers significance. In particular, how significant is the the discovery of a moderately sized submatrix of 1s in a large data matrix? To address this question, we establish probability bounds on the size of the largest submatrix of 1s in a random binary matrix. These bounds improve upon existing inequalities in the literature, and yield approximate p-values for discovered submatrices under the null hypothesis that the data consists of independent Bernoulli random variables.

Much of the data to which data mining methods are applied are obtained by high-throughput technologies or the automated collection of data from diverse sources with varying levels of reliability. The resulting data sets are often subject to moderate levels of error and noise. Our second question involves the behavior and performance of FIM in the presence of noise. Standard frequent itemset algorithms do not account for noise or errors in their search for submatrices of 1s. We consider the noise sensitivity of FIM under a simple binary additive noise model and show that, even at small noise levels, blocks of 1s is broken into fragments of logarithmic size. Thus such blocks cannot be directly recovered by standard frequent itemset algorithms.

Lastly, we consider the problem of recovering a block of 1s in the presence of additive noise using an error-tolerant criterion (approximate frequent itemsets) that allows submatrices containing a limited fraction of zeros. We show how the AFI criterion can recover a square submatrix of 1s against a background of 0s, even when the size of the target submatrix is very small.

1.1 Overview

The next section contains several results on the size of maximal submatrices of ones in a random matrix with independent Bernoulli entries. In addition, we present a small simulation study that explores the applicability of the asymptotic theory to small samples. Section 2 is devoted to the description of the additive noise model and the noise sensitivity of standard FIM. In Section 3 we consider the recoverability of block structures in the presence of noise using the approximate frequent itemset criterion.

2 Frequent Itemsets in Random Matrices

There is a large literature on ordinary (full-space) clustering that spans 50 years. While there has been recent attention and progress on the problem of cluster validation [30], [17], [11], there is no general and systematic treatment of signifiance for full-space clustering. FIM is more amenable to significance analysis

than full-space clustering, as attention shifts from high-dimensional objects (the rows or columns of the data matrix) to the entries of the data matrix itself, which are organized into a simple two-dimensional array. Here we consider several questions related to the size and statistical significance of frequent itemsets in a random matrix. The focus is on the size of the largest submatrix of 1s, or a specified fraction of 1s, in a binary matrix with Bernoulli entries. For simplicity of exposition, we emphasize the case of square matrices and square submatrices. Some extensions to the non-square case are described in Section 2.3 below.

2.1 Square Submatrices of 1s

Let $\mathbf{X}$ be an $m \times n$ binary matrix. A submatrix of $\mathbf{X}$ is a collection $\mathbf{U} = \{x_{i,j} : i \in A, j \in B\}$ where $A \subseteq \{1, \ldots, m\}$ and $B \subseteq \{1, \ldots, n\}$. The Cartesian product $C = A \times B$ will be called the index set of $\mathbf{U}$. Given $C = A \times B$, define $\mathbf{X}[C]$ to be the submatrix of $\mathbf{X}$ with index set C. When no ambiguity will arise, C will also be referred to as a submatrix of $\mathbf{X}$. Note that $\mathbf{X}$ can be viewed as the adjacency matrix of a bi-partite graph $G(\mathbf{X})$. The graph $G(\mathbf{X})$ has vertex set V equal to the disjoint union $V = V_r \cup V_c$, where V_r corresponds to the rows of $\mathbf{X}$, V_c corresponds to its columns, and there is an edge between $i \in V_r$ and $j \in V_c$ if and only if $x_{i,j} = 1$. With this association, submatrices of ones in $\mathbf{X}$ are in one-to-one correspondence with bi-cliques in $G(\mathbf{W})$. This connection is the basis for the SAMBA bi-cluster algorithm of Tanay *et al.* [29].

Definition: Given any binary matrix $\mathbf{X}$, let $M(\mathbf{X})$ be the largest value of k such that $\mathbf{X}$ contains a $k \times k$ submatrix of 1s.

Definition: Let $\mathbf{Z}_n$ denote an $n \times n$ binary matrix whose entries are independent Bernoulli(p) random variables, with $p \in (0, 1)$. We will write $\mathbf{Z}_n \sim \text{Bern}(p)$.

A natural starting point for studying the significance of FIM is $M(\mathbf{Z}_n)$, the size of the largest submatrix of 1s in a binary matrix with independent Bernoulli entries. To obtain bounds on $M(\mathbf{Z}_n)$, let $U_k(n)$ be the number of $k \times k$ submatrices of 1s in $\mathbf{Z}_n$. Then, using Stirling's approximation, it is easy to show that $EU_k(n) = \binom{n}{k}^2 p^{k^2} \approx (2\pi)^{-1} n^{2n+1} k^{-2k-1} (n-k)^{-2(n-k)-1} p^{k^2}$. The first moment method from combinatorial probability suggests that $M(\mathbf{Z}_n)$ will be close to the value of k for which $EU_k(n) = 1$. Accordingly, define $s(n)$ to be any positive solution of the equation

$$1 = \phi_n(s) = (2\pi)^{-\frac{1}{2}} n^{n+\frac{1}{2}} s^{-s-\frac{1}{2}} (n-s)^{-(n-s)-\frac{1}{2}} p^{\frac{s^2}{2}}. \tag{1}$$

A routine but involved analysis shows that any solution $s(n)$ of (1) must satisfy the relation

$$s(n) = 2 \log_b n - 2 \log_b \log_b n + C + o(1), \tag{2}$$

where $b = p^{-1}$ and C is a positive constant. Moreover, by standard calculus, it can be shown that $\phi_n(\cdot)$ is monotone decreasing when $\log_b n < s < 2 \log_b n$. Thus when n is sufficiently large, there is only one solution of (1) in the interval

$(\log_b n, 2\log_b n)$, and therefore $s(n)$ is uniquely defined. Define $k(n) = \lceil s(n) \rceil$. A simple application of the first moment method yields a bound on the probability that $M(\mathbf{Z}_n)$ is larger than $k(n)$, which can be used to assess the statistical significance of submatrices of ones identified by bi-clustering algorithms.

Proposition 1 *Fix $0 < \gamma < 1$. When n is sufficiently large, for every integer $1 \le r \le \gamma\, n$ we have $P\{M(\mathbf{Z}_n) \ge k(n) + r\} \le 2\, n^{-2\,r}\, (\log_b n)^{3r}$, where $b = p^{-1}$.*

Proof: To establish the bound with n independent of r, it suffices to consider a sequence $r = r_n$ that changes with n in such a way that $1 \le r_n \le \gamma\, n$. Fix n for the moment, let $l = k(n) + r_n$, and let $U_l(n)$ be be the number of $l \times l$ submatrices of 1s in $\mathbf{Z}_n$. Then by Markov's inequality and Stirling's approximation,

$$P(M(\mathbf{Z}_n) \ge r) \;=\; P(U_l \ge 1) \;\le\; E(U_l) \;=\; \binom{n}{l}^2 p^{l^2} \;\le\; 2\phi_n(l)^2. \qquad (3)$$

A straightforward calculation using the definition of $\phi_n(\cdot)$ shows that

$$2\phi_n(l)^2 \;=\; 2\,\phi_n^2(k(n))\, p^{r\cdot k(n)}\, [\, A_n(r)\, B_n(r)\, C_n(r)\, D_n(r)\,]^2, \qquad (4)$$

where

$$A_n(r) = \left(\frac{n-r-k(n)}{n-k(n)}\right)^{-n+r+k(n)+\frac{1}{2}} \qquad B_n(r) \;=\; \left(\frac{r+k(n)}{k(n)}\right)^{-k(n)-\frac{1}{2}}$$

$$C_n(r) = \left(\frac{n-k(n)}{r+k(n)}\, p^{\frac{k(n)}{2}}\right)^{r} \qquad D_n(r) \;=\; p^{\frac{r^2}{2}}$$

Note that $p^{r\cdot k(n)} = o(n^{-2r}(\log_b n)^{3r})$, and that $\phi_n^2(k(n)) \le 1$ by the monotonicity of $\phi_n(\cdot)$ and the definition of $k(n)$. Thus it suffices to show that $A_n(r) \cdot B_n(r) \cdot C_n(r) \cdot D_n(r) \le 1$ when n is sufficiently large. To begin, note that for any fixed $\delta \in (0, 1/2)$, when n is sufficiently large,

$$C_n(r)^{\frac{1}{r}} \;=\; \frac{n-k(n)}{r+k(n)}\, p^{\frac{k(n)}{2}} \;\le\; \frac{n}{k(n)}\, p^{\frac{k(n)}{2}} \;\le\; \frac{n}{(2-\delta)\log_b n}\, \frac{\frac{2+\delta}{2}\log_b n}{n}$$

which is less than one. In order to show $A_n(r) \cdot B_n(r) \cdot D_n(r) \le 1$, we consider two possibilities for the asymptotic behavior of $r = r_n$.

Case 1: Suppose $r/k(n) \to 0$ as $n \to \infty$. In this case, $B_n(r)^{\frac{1}{r}} = (1 + o(1))\, e^{-1}$. Moreover, $r/n \to 0$, which implies that $A_n(r)^{\frac{1}{r}} = (1 + o(1))\, e$. Thus

$$A_n(r) \cdot B_n(r) \cdot D_n(r) \;=\; ((1 + o(1))^2\, p^{\frac{r}{2}})^r \;\le\; 1$$

when n is sufficiently large.

Case 2: Suppose $\liminf_n r/k(n) > 0$. In this case a routine calculation shows that $B_n(r) \le 1$ for any $r \ge 1$, so it suffices to show that

$$A_n(r) \cdot D_n(r) \;\le\; 1. \qquad (5)$$

Note that $D_n(r) = (p^{\frac{r}{2}})^r$ and $A_n(r)^{\frac{1}{r}} = (1+o(1))\,e$ when $r = o(n-k(n))$. Thus (5) holds when $r = o(n-k(n))$.

It remains to consider the case $o(n-k(n)) < r < \gamma\, n$. As $\sqrt{(2+\frac{2}{1-\gamma})\, n/\log b} = o(n-k(n))$, it suffices to assume that $\sqrt{(2+\frac{2}{1-\gamma})\, n/\log b} < r < \gamma\, n$. In this case,

$$\begin{aligned}\log_b A_n(r)\cdot D_n(r) \;&=\; \log_b\left[\left(1+\frac{r}{n-k(n)-r}\right)^{n-r-k(n)-\frac{1}{2}} p^{\frac{r^2}{2}}\right]\\ &\le\; n\log_b\left(1+\frac{r}{n-r-k(n)}\right) - \frac{(2+\frac{2}{1-\gamma})\, n}{2\log b} \;\le\; 0,\end{aligned}$$

where the last inequality comes from the fact that $\log_b(1+x) \le x/\log b$ for $x \ge 0$. ∎

An inspection of the proof shows that the inequality of Proposition 1 is obtained via a standard union (Bonferroni) type bound on the probability of finding a $k\times k$ submatrix of 1s in $\mathbf{Z}_n$. In general, union bounds are rather loose, and indeed, with additional calculation, one can improve the upper bound in Proposition 1 to $n^{-(4-\delta)\,r}\,(\log_b n)^{3r}$ for any $\delta > 0$. Nevertheless, a more refined second moment argument (see Theorem 1 below) shows that the threshold $k(n)$ can not be improved.

Bollobás [5] and Grimmett and McDiarmid [12] established analogous bounds for the size of a maximal clique in a random graph, with the larger threshold $k(n) = 2\log_b n$. Koyutürk and Szpankowski [16] studied the problem of finding dense patterns in binary data matrices. They used a Chernoff type bound for the binomial distribution to assess whether an individual submatrix has an enriched fraction of ones, and employed the resulting test as the basis for a heuristic search for significant bi-clusters. Tanay *et al.* [28] assessed the significance of bi-clusters in a real-valued matrix using likelihood-based weights, a normal approximation and a standard Bonferroni bound to account for the multiplicity of submatrices.

As noted above, $M(Z_n)$ is the size of the largest bi-clique in a random $n\times n$ bi-partite graph. Bollobás and Erdős [4] and Matula [20] studied the size of the largest clique in a standard random graph with n vertices, where each edge is included with probability p, independent of the other edges. In particular, they obtained strong almost sure results on the asymptotic size of maximal cliques. Bollobás [5] gives a good account of these results. By extending the arguments in [4, 20] to bi-cliques one may establish the following analogous result; the proof is rather technical and is omitted. Assume that for each n the matrix $\mathbf{Z}_n$ is the upper left corner of an infinite array $\{z_{i,j} : i,j \ge 1\}$ of Bernoulli(p) random variables with $0 < p < 1$.

Theorem 1 *With probability one, $|M(\mathbf{Z}_n) - s(n)| < \frac{3}{2}$ when n is sufficiently large. Thus $M(\mathbf{Z}_n)$ eventually concentrates on one of the (at most three) integers within distance $3/2$ of the real number $s(n)$.*

Dawande *et al.* [8] used first and second moment arguments to show (in our terminology) that $P(\log_b n \leq M(\mathbf{Z}_n) \leq 2\log_b n) \to 1$ as n tends to infinity. Extending this work, Park and Szpankowski [25] showed that if $\tilde{M}$ is the side-length of the largest square submatrix of 1s in an $m \times n$ Bernoulli matrix, then $P(|\tilde{M} - \log_b(mn)| > \epsilon \log(mn)) \leq O((mn)^{-1}(\log(mn))^6)$. When $m = n$ their result implies that $(2-\epsilon)\log_b n \leq M(\mathbf{Z}_n) \leq (2+\epsilon)\log_b n$ eventually almost surely.

2.2 Submatrices with Large Fraction of Ones

In situations where noise is present, one may wish to look for submatrices having a large fraction of 1s, rather than requiring the stronger condition that every entry be equal to 1. Let $\mathbf{X}$ be a binary matrix, and let $\mathbf{U}$ be a submatrix of $\mathbf{X}$ with index set C. Let

$$F(\mathbf{U}) \;=\; |C|^{-1} \textstyle\sum_{(i,j)\in C} x_{i,j}$$

be the fraction of ones in $\mathbf{U}$. Fix $\tau \in (0,1)$ and define $M_\tau(\mathbf{X})$ to be the largest k such that $\mathbf{X}$ contains a $k \times k$ submatrix $\mathbf{U}$ with $F(\mathbf{U}) \geq \tau$.

Proposition 2 *Fix $0 < \gamma < 1$ and suppose that $0 < p < \tau < 1$. When n is sufficiently large, $P(M_\tau(\mathbf{Z}_n) \geq 2\log_{b^*} n + r) \leq 2n^{-2r}(\log_{b^*} n)^{3r}$ for each $1 \leq r \leq \gamma n$. Here $b^* = \exp\{3(\tau-p)^2/8p\}$.*

Proof: For $l \geq 1$ let $V_l(n)$ be the number of $l \times l$ submatrices $\mathbf{U}$ of $\mathbf{Z}_n$ with $F(\mathbf{U}) \geq \tau$. Note that $E(V_l(n)) = \binom{n}{l}^2 P\left(F(\mathbf{Z}_l) \geq \tau\right)$. The random variable $l^2 \cdot F(\mathbf{Z}_l)$ has a Binomial(l^2, p) distribution. Using a standard inequality for the tails of the binomial distribution, (*c.f.* Problem 8.3 of [9]), we find that $P\left(F(\mathbf{Z}_l) \geq \tau\right) \leq q^{l^2}$ where $q = 1/b^*$. It then follows from Stirling's approximation that $EV_l(n) \leq 2$ when $l = l(n) = 2\log_{b^*} n$. For $l = r + l(n)$, $P(M_\tau(\mathbf{Z}_n) \geq l) \leq E(V_l(n))$ and the stated inequality then follows from arguments analogous to those in the proof of Proposition 1. ■

Note that the base $b^* = \exp\{3(\tau-p)^2/8p\}$ may not always yield the best upper bound. When $p \geq \frac{1}{2}$, b^* can be replaced by $\exp\{(\tau-p)^2/2p(1-p)\}$ (*cf.* [22]). When $\tau \to 1$, $b^* = \exp\{3(\tau-p)^2/8p\}$ fails to converge to p^{-1}, so that the probability bound above does not coincide with that of Proposition 1. In this case, the disparity may be remedied by using an alternative bound for the tails of the binomial (*e.g.* [15]) and a corresponding base for the logarithm.

2.3 Non-square Matrices

The restriction to square matrices above can readily be relaxed, yielding bounds for data sets with more transactions than items, or vice versa. Suppose that $\mathbf{Z}_{m,n} \sim \text{Bern}(p)$ is an $m \times n$ random matrix with $\frac{m}{n} = \alpha$ for some $\alpha > 0$. For any $\rho \geq 1$, let $M_\alpha^\rho(\mathbf{Z})$ be the largest k such that there exists at least one $\lceil \rho k \rceil \times k$ submatrix of 1s in $\mathbf{Z}$. One may extend Proposition 1 as follows.

Proposition 3 *Fix $0 < \gamma < 1$. When n is sufficiently large,*

$$P\{M_\alpha^\rho(\mathbf{Z}) \;\geq\; k(\alpha,\rho,n)+r\} \;\leq\; n^{-(\rho+1)\,r}\,2(\log_b n)^{(\rho+2)r} \tag{6}$$

for each $1 \leq r \leq \gamma\, n$. Here $k(\alpha,\rho,n) = \frac{\rho+1}{\rho}\log_b n + \log_b \frac{\alpha}{\rho}$.

One may generalize Proposition 3 to submatrices with large fractions of 1s by replacing b with the base b^* of Proposition 2.

Park and Szpankowski [25] established probability bounds on the maximum area of non-square submatrices and showed that such submatrices have aspect ratio close to zero. In addition they established probability bounds for square submatrices (discussed above) that provide weaker inequalities like those in Proposition 3 when $\rho = 1$.

Propositions 1, 2 and 3 can provide bounds on the statistical significance of submatrices discovered by frequent itemset mining algorithms, under the null hypothesis that the observed data is purely random. Suppose for example that an FIM algorithm is applied to a $4,000 \times 100$ binary matrix $\mathbf{Y}$, 65% of whose entries are equal to 1. Suppose that the algorithm finds a 44×25 submatrix $\mathbf{U}$ of ones in $\mathbf{Y}$. Applying Proposition 3 with $p = 0.65$, $\alpha = 40$ and $\rho = 1.76$ we find that $k(\alpha,\rho,n) = 24$ and that the probability of finding such a matrix $\mathbf{U}$ in a purely random matrix is at most

$$2\,n^{-(1.76+1)\times(25-24)}\,(\log_b n)^{(1.76+2)\times(25-24)} \;\approx 0.04467.$$

Thus $\mathbf{U}$ may be assigned a p-value $p(\mathbf{U}) \leq 0.04467$. On the other hand, consider the case that an error tolerant FIM algorithm finds an 73×25 submatrix $\mathbf{U}'$ in Y with 95% 1s. Since in this case $p > \frac{1}{2}$, the discussion immediately after Proposition 2 suggests using $b^* = \exp\{(0.95-p)^2/2p(1-p)\} = 1.2187$ for a better bound. By plugging each corresponding term into (6), one obtains a nominal p-value $p(\mathbf{U}') \leq 0.04802$.

2.4 Simulations

The results of the previous section hold for n sufficiently large. To test their validity for moderate values of n, we carried out a simple simulation study on Z_n with $n = 40$ and 80, and $p = .2$. In each case, we generated 400 such matrices and applied the FP-growth algorithm [13] to identify all maximal submatrices of ones. For each maximal submatrix of ones we recorded the length of its shorter side. The maximum of these values is $M(\mathbf{Z}_n)$. We recorded $M(\mathbf{Z}_n)$ in each of the 400 simulations and compared its value to the corresponding bounds $s(40) \approx 3.553$ and $s(80) \approx 4.582$. Table 1 summarizes the results. In each case $|M(\mathbf{Z}_n) - s(n)| \leq 1$.

3 Noise Sensitivity of FIM

3.1 Statistical Noise Model

In order to account for and study the potential effects of noise on FIM, we consider a simple noise model. Under the model the observed data matrix $\mathbf{Y}$ is

Table 1. Simulation results on $\hat{M}(Z_n)$ based on 400 replications for each n.

n	$s(n)$	k	Proportion of $M(\mathbf{Z}_n) = k$
40	3.553	3	85.75%
		4	14.25%
80	4.582	4	97%
		5	3%

equal to the component-wise modulo 2 sum of a "true" unobserved data matrix $\mathbf{X}$ and a noise matrix $\mathbf{Z}$ whose entries are independent Bernoulli(p) random variables. Formally,

$$\mathbf{Y} \;=\; \mathbf{X} \oplus \mathbf{Z} \tag{7}$$

so that $y_{i,j} = x_{i,j}$ if $z_{i,j} = 0$ and $y_{i,j} = 1 - x_{i,j}$ if $z_{i,j} = 1$. The model (7) is the binary version of the standard additive noise model in statistical inference. It is equivalent to the simple communication model, widely studied in information theory, in which the values of $\mathbf{X}$ are observed after being passed through a memoryless binary symmetric channel.

3.2 Noise Sensitivity

If the matrix $\mathbf{X}$ in (7) contains interesting structure, for example a large submatrix of ones, there is reason to hope that this structure would be readily apparent in the observed matrix $\mathbf{Y}$ and could be recovered by standard frequent itemset algorithms without much effort. Unfortunately this is not necessarily the case, as the next result shows.

Let $\mathbf{X}$ be an $n \times n$ binary matrix, and let $\mathbf{Y} = \mathbf{X} \oplus \mathbf{Z}$ with $\mathbf{Z} \sim \text{Bern}(p)$ and $0 < p < \frac{1}{2}$. We are interested in how $M(\mathbf{Y})$ depends on $\mathbf{X}$, and in particular how the value of $M(\mathbf{Y})$ reflects block structures (submatrices of 1s) in $\mathbf{X}$. If $\mathbf{X} = \mathbf{0}$ then $\mathbf{Y} \sim \text{Bern}(p)$. In this case, Proposition 1 and Theorem 1 ensure that $M(\mathbf{Y})$ is roughly $2 \log_b n$ with $b = p^{-1}$. At the other extreme, if $\mathbf{X} = \mathbf{1}$ then it is easy to see that $\mathbf{Y} \sim \text{Bern}(1-p)$, and in this case $M(\mathbf{Y})$ is roughly $2 \log_{b'} n$ with $b' = (1-p)^{-1}$. The latter case represents the best possible situation in regards to maximizing $M(\mathbf{Y})$.

Proposition 4 *Let $b' = (1-p)^{-1}$ and fix $0 < \gamma < 1$. When n is sufficiently large, $P\{M(\mathbf{Y}) \geq 2 \log_{b'} n + r\} \leq 2\, n^{-2\, r}\, (\log_{b'} n)^{3r}$ for every matrix X and for every integer $1 \leq r \leq \gamma\, n$.*

Proof: Fix n and let $\mathbf{W}_n = \{w_{i,j}\}$ be an $n \times n$ binary matrix with independent entries, defined on the same probability space as $\{z_{i,j}\}$, such that

$$w_{i,j} \;=\; \begin{cases} \text{Bern}\left(\frac{1-2p}{1-p}\right) & \text{if } \; x_{ij} = y_{ij} = 0 \\ 1 & \text{if } \; x_{ij} = 0, y_{ij} = 1 \\ y_{i,j} & \text{if } \; x_{ij} = 1 \end{cases} \tag{8}$$

Note that the above definition is valid since we assume $p < \frac{1}{2}$ here. Define $\tilde{\mathbf{Y}}_n = \mathbf{Y}_n \vee \mathbf{W}_n$ to be the entrywise maximum of $\mathbf{Y}_n$ and $\mathbf{W}_n$. Clearly $M(\mathbf{Y}_n) \leq M(\tilde{\mathbf{Y}}_n)$, as any submatrix of ones in $\mathbf{Y}_n$ must also be present in $\tilde{\mathbf{Y}}_n$. Moreover, it is easy to check that $P(\tilde{y}_{i,j} = 1) = 1 - p$ for each $1 \leq i, j \leq n$, so that $\tilde{\mathbf{Y}}_n \sim \text{Bern}(1-p)$. The result now follows from Proposition 1. ∎

Proposition 4 has the following consequence. No matter what type of block structures might exist in $\mathbf{X}$, in the presence of random noise these structures leave behind only logarithmic sized fragments in the observed data matrix. In particular, under the additive noise model (7) block structures in $\mathbf{X}$ cannot be recovered, even approximately, by standard frequent itemset algorithms that look for submatrices of ones without errors.

4 Recovery

Here we consider the simple problem of recovering, in the presence of noise, a submatrix of ones against a background of zeros. Proposition 4 shows that standard FIM algorithms are sensitive to noise, and are not readily applicable to the recovery problem. This shortcoming can be remedied by algorithms that look instead for submatrices having a large *fraction* of ones. Several recent papers [24, 18, 19, 27, 23, 6, 31] in the data mining literature have addressed this question, each using a criterion that weakens the all 1s model of FIM. Below we show how one such criterion, introduced in [18], can be used to recover block structures in noise.

Let $\mathbf{X}$ be an $n \times n$ binary matrix that consists of an $l \times l$ submatrix of ones, with index set C^*, and all other entries equal to 0. (The rows and columns of C^* need not be contiguous.) Given an observation $\mathbf{Y} = \mathbf{X} \oplus \mathbf{Z}$ of $\mathbf{X}$ with $\mathbf{Z} \sim \text{Bern}(p)$ and $0 < p < 1/2$, we wish to recover the submatrix C^*.

Let p_0 be any number such that $p < p_0 < 1/2$, and let $\tau = 1 - p_0$ be an associated error threshold. If $\mathbf{U}$ is an $a \times b$ submatrix of $\mathbf{Y}$, denote its rows and columns by $u_{1*}, \ldots, u_{a*}$ and $u_{*1}, \ldots, u_{*b}$, respectively. The following definition of error-tolerant itemsets was introduced in [18]. An algorithm for finding such itemsets is given in [19].

Definition: An $a \times b$ submatrix C of $\mathbf{Y}$ is a *τ-approximate frequent itemset* (AFI) if $F(u_{i*}) \geq \tau$ and $F(u_{*j}) \geq \tau$ for each $i = 1, \ldots, a$ and $j = 1, \ldots, b$. Let $\text{AFI}_\tau(\mathbf{Y})$ be the collection of all τ-AFIs in $\mathbf{Y}$.

We estimate C^* by the index of the largest square AFI in the observed matrix $\mathbf{Y}$. More precisely, let $\mathcal{C}$ be the family of index sets of square submatrices $C \in \text{AFI}_\tau(\mathbf{Y})$, and define

$$\hat{C} \;=\; \text{argmax}_{C \in \mathcal{C}} |C|$$

to be any maximal sized submatrix in $\mathcal{C}$. Note that $\mathcal{C}$ and $\hat{C}$ depend only on the observed matrix $\mathbf{Y}$. Let the ratio

$$\Lambda \;=\; |\hat{C} \cap C^*|/|\hat{C} \cup C^*|$$

measure the overlap between the estimated index set $\hat{C}$ and the true index set C^*. Thus $0 \leq \Lambda \leq 1$, and values of Λ close to one indicate better overlap.

Theorem 2 *Let $0 < p < p_0 < 1/2$ and $\tau = 1 - p_0$. When n is sufficiently large, for any $0 < \alpha < 1$ such that $12\alpha^{-1}(\log_b n + 2) \leq l$ we have*

$$P\left(\Lambda \leq \frac{1-\alpha}{1+\alpha}\right) \;\leq\; \Delta_1(l) + \Delta_2(\alpha, l).$$

Here $\Delta_1(l) = 2e^{-\frac{l(p-p_0)^2}{3p}}$, $\Delta_2(\alpha, l) = 2n^{-\frac{1}{6}\alpha l + 2\log_b n}$, and $b = \exp\{3(1-2\,p_0)^2/8p\}$.

The conditions of Theorem 2 require that the noise level $p < 1/2$ and that the user-specified parameter p_0 satisfies $p < p_0 < 1/2$. Thus, in advance, one only needs to know an upper bound on the noise level p. Theorem 2 can readily be applied to the asymptotic recovery of structure in a sequential framework. Suppose that $\{\mathbf{X}_n : n \geq 1\}$ is a sequence of square binary matrices, where $\mathbf{X}_n$ is $n \times n$ and consists of an $l_n \times l_n$ submatrix C_n^* of 1s with all other entries equal to 0. For each n we observe $\mathbf{Y}_n = \mathbf{X}_n \oplus \mathbf{Z}_n$, where $\mathbf{Z}_n \sim$Bern(p). Let Λ_n measure the overlap between C_n^* and the estimate $\hat{C}_n$ produced by the AFI recovery method above. The following result follows from Theorem 2 and the Borel Cantelli lemma.

Corollary 1 *If $l_n \geq 12\psi(n)(\log_b n + 2)$ where $\psi(n) \to \infty$ as $n \to \infty$, then eventually almost surely*

$$\Lambda_n \;\leq\; \frac{1-\psi(n)^{-1}}{1+\psi(n)^{-1}}.$$

Reuning-Scherer studied several recovery problems in [26]. In the case considered above, he calculated the fraction of 1s in every row and every column of $\mathbf{Y}$, and then selected those rows and columns with a large fraction of 1s. His algorithm is consistent when $l \geq n^\alpha$ for $\alpha > 1/2$. However, a simple calculation using the central limit theorem demonstrates that individual row and column sums alone are not sufficient to recover C^* when $l \leq n^\alpha$ for $\alpha < 1/2$. In this case, one gains considerable power by directly considering submatrices and, as the result above demonstrates, one can consistently recover C_n^* if $l_n/\log n \to \infty$.

The following two lemmas will be used in the proof of Theorem 2. Lemma 1 implies that $|\hat{C}|$ is greater than or equal to $|C^*|$ with high probability. Lemma 2 shows that $\hat{C}$ can only contain a small proportion of entries outside C^*. The proofs of Lemma 1, and a sketch of the proof of Lemma 2, can be found in the Appendix.

Lemma 1 *Under the conditions of Theorem 2, $P\left(|\hat{C}| < l^2\right) \leq \Delta_1(l)$.*

Lemma 2 *Let $\mathcal{A}$ be the collection of $C \in \mathcal{C}$ such that $|C| > \frac{l^2}{2}$ and $\frac{|C \cap C^{*c}|}{|C|} \geq \alpha$. Let A be the event that $\mathcal{A} \neq \emptyset$. If n is sufficiently large, then $l \geq 12\alpha^{-1}(\log_b n + 2)$ implies*

$$P(A) \ \leq \ \Delta_2(\alpha, l)$$

Proof of Theorem 2: Let E be the event that $\{\Lambda \leq \frac{1-\alpha}{1+\alpha}\}$. It is clear that E can be expressed as the union of two disjoint events E_1 and E_2, where

$$E_1 = \{|\hat{C}| < |C^*|\} \cap E \tag{9}$$

and

$$E_2 = \{|\hat{C}| \geq |C^*|\} \cap E \tag{10}$$

One can bound $P(E_1)$ by $\Delta_1(l)$ via Lemma 1.

It remains to bound $P(E_2)$. By the definition of Λ, the inequality $\Lambda \leq \frac{1-\alpha}{1+\alpha}$ can be rewritten equivalently as

$$1 + \frac{|\hat{C} \cap C^{*c}|}{|\hat{C} \cap C^*|} + \frac{|\hat{C}^c \cap C^*|}{|\hat{C} \cap C^*|} \geq \frac{1+\alpha}{1-\alpha}.$$

When $|\hat{C}| \geq |C^*|$, one can verify that $|\hat{C} \cap C^{*c}| \geq |\hat{C}^c \cap C^*|$, which implies that

$$1 + \frac{|\hat{C} \cap C^{*c}|}{|\hat{C} \cap C^*|} + \frac{|\hat{C}^c \cap C^*|}{|\hat{C} \cap C|} \leq 1 + 2\frac{|\hat{C} \cap C^{*c}|}{|\hat{C} \cap C^*|}.$$

Therefore, $E_2 \subset E_2^*$, where

$$\begin{aligned} E_2^* &= \{|\hat{C}| \geq |C^*|\} \cap \left\{1 + 2\frac{|\hat{C} \cap C^{*c}|}{|\hat{C} \cap C^*|} \geq \frac{1+\alpha}{1-\alpha}\right\} \\ &\subset \{|\hat{C}| > \frac{l^2}{2}\} \cap \left\{1 + 2\frac{|\hat{C} \cap C^{*c}|}{|\hat{C} \cap C^*|} \geq \frac{1+\alpha}{1-\alpha}\right\}. \end{aligned}$$

Notice that $1 + 2\frac{|\hat{C} \cap C^{*c}|}{|\hat{C} \cap C^*|} \geq \frac{1+\alpha}{1-\alpha}$ implies $\frac{|\hat{C} \cap C^{*c}|}{|\hat{C}|} \geq \alpha$. Therefore, by Lemma 2, $P(E_2^*) \leq \Delta_2(\alpha, l)$. ∎

5 Conclusion

The problem of data mining has commonly been approached from the point of view of data structures and algorithms, in a setting that is primarily deterministic. This paper addresses several statistical questions related to the basic problem of frequent itemset mining, namely significance, noise-tolerance and recovery. The probabilistic bounds given here provide a preliminary basis for assessing the significance of discovered itemsets, with or without errors, and give one objective criterion for sifting through the (potentially large) number of frequent itemsets in a data matrix. The results on the noise sensitivity of standard FIM provide some justification for the current efforts on error-tolerant algorithms. Further justification is provided by the use of one such method for recovery of block structures.

6 Acknowledgments

The research presented here was supported part by NSF grant DMS 040636.

7 Appendix

Proof of Lemma 1: Let $u_{1*}, ..., u_{l*}$ be corresponding rows of C^* in $\mathbf{Y}$ and let V be the number of rows satisfying $F(u_{i*}) < 1 - p_0$, where $F(\cdot)$ is the function measuring the fraction of ones. By Markov's inequality,

$$P(V \geq 1) \ \leq \ E(V) \ = \ \sum_{i=1}^{l} P(F(u_{i*}) < 1 - p_0). \tag{11}$$

Using standard bounds on the tails of the binomial distribution, when l_n is sufficiently large,

$$P(V \geq 1) \leq l \cdot e^{-\frac{3l(p-p_0)^2}{8p}} \leq e^{-\frac{1}{3p}l(p-p_0)^2}, \tag{12}$$

when l is sufficiently large.

Let $u_{*1}, ..., u_{*l}$ be corresponding columns of C^* in $\mathbf{Y}$ and let V' be the number of columns satisfying $F(u_{*i}) < 1 - p_0$. A similar calculation as above shows that

$$\begin{aligned} P(V' \geq 1) \leq E(V') \ &\leq \ l \cdot e^{-3\frac{l(p-p_0)^2}{8p}} \\ &\leq \ e^{-\frac{1}{3p}l(p-p_0)^2}. \end{aligned}$$

Since $\{|\hat{C}| < l^2 = |C^*|\} \subset \{C^* \notin AFI_\tau(\mathbf{Y})\} \ \subset \ \{V \geq 1\} \cup \{V' \geq 1\}$,

$$\begin{aligned} P\{|\hat{C}| < l^2\} \ &\leq \ P(V \geq 1) + P(V' \geq 1) \\ &\leq 2e^{-\frac{1}{3p}l_n(p-p_0)^2} = \Delta_1(l). \quad \blacksquare \end{aligned}$$

The proof of Lemma 2, relies on two basic facts below. The proof of Fact 2 is technical and is omitted.

Fact 1: Given $0 < \tau_0 < 1$, if there exists a $k \times r$ binary matrix M satisfying $F(M) \geq \tau_0$, then for $v = \min\{k, r\}$, there exists a $v \times v$ submatrix D of M such that $F(D) \geq \tau_0$.

Proof: Without loss of generality, we assume $v = k \leq r$. Then we rank each column according to its fraction of ones, and reorder the columns in descending order. Let the reordered matrix be M^1. Let $D = M^1[(1, ..., v) \times (1, ..., v)]$. One can verify that $F(D) \geq \tau_0$. ■

Fact 2: Let $1 < \gamma < 2$ be a constant, and let W be an $n \times n$ binary matrix. Let R_1 and R_2 be two square submatrices of W satisfying (i) $|R_2| = k^2$, (ii) $|R_1 \backslash R_2| > k^\gamma$ and (iii) $R_1 \in \mathrm{AFI}_\tau(W)$. Then there exists a square submatrix $D \subset R_1 \backslash R_2$ such that $|D| \geq k^{2\gamma-2}/9$ and $F(D) \geq \tau$.

Proof of Lemma 2: If $C \in \mathcal{A}$ then
(i) $|C^*| = l^2$,
(ii) $|C \backslash C^*| = |C| \cdot \frac{|C \cap C^{*c}|}{|C|} \geq \frac{l^2 \cdot \alpha}{2} = l^\gamma$, where $\gamma = 2 + \log_l \frac{\alpha}{2}$,
(iii) $C \in \mathrm{AFI}_{1-p_0}(\mathbf{Y})$.
Thus, by Fact 2, there exists a $v \times v$ submatrix D of $C \backslash C^*$ such that $F(D) \geq 1 - p_0$ and $v \geq \frac{\alpha l}{6}$, which implies that

$$\max_{c \in \mathcal{C}} M^\tau(C \cap C^{*c}) \geq v \geq \frac{\alpha l}{6},$$

where $\tau = 1 - p_0$.

Let $\mathbf{W}(\mathbf{Y}, C^*)$ be a $n \times n$ binary random matrix, where $w_{ij} = y_{ij}$ if $(i,j) \notin C^*$, and $w_{ij} \sim Bern(p)$ otherwise. It is clear that

$$M^\tau(\mathbf{W}) \geq \max_{c \in \mathcal{C}} M^\tau(C \cap C^{*c}) \geq \frac{\alpha l}{6}.$$

By Proposition 2, when n is sufficiently large and $l \geq 12\alpha^{-1}(\log_b n + 2)$, we can bound $P(A)$ with

$$\begin{aligned} P(A) &\leq P(\max_{c \in \mathcal{C}} M^\tau(C \cap C^{*c}) \geq \frac{\alpha l}{6}) \\ &\leq P(M^\tau(\mathbf{W}) \geq \frac{\alpha l}{6}) \leq 2n^{-(\alpha l/6 - 2\log_{b'} n)}, \end{aligned} \tag{13}$$

where $b' = e^{\frac{3(1-p_0-p)^2}{8p}}$. As $p_0 > p$, it is trivial to verify that $b < b'$. Consequently, one can bound the RHS of inequality (13) by $\Delta_2(\alpha, l)$. ∎

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. *Proceedings of ACM SIGMOD'93*, 207 - 216, 1993.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A Verkamo. Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, 307 - 328, AAAI/MITPress, 1996.

[3] R. Agrawal, J. Gehrke, D. Gunopulos and P.Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. *Proceedings of ACM SIGMOD'98*, 94 - 105, 1998.

[4] B. Bollobás and P. Erdős. Cliques in random graphs. *Math. Proc. Cam. Phil. Soc.*, 80: p 419 - 427, 1976.

[5] B. Bollobás. *Random Graphs* second edition. Cambridge Studies in Advanced Mathematics, 2001.

[6] D. Chakrabarti, S. Papadimitriou, D. Modha and C. Faloutsos. Fully Automatic Cross-Associations. *Proceedings of ACM SIGKDD'04*, 79 - 88, 2004.

[7] Y. Cheng and G.M. Church. Biclustering of expression data. *Proceedings of ISMB'00*, 93- 103, 2000.

[8] M. Dawande, P. Keskinocak, J. Swaminathan, and S. Tayur. On bipartite and multipartite clique problems. *J. Algorithms* 41(2): 388 - 403, 2001.

[9] L. Devroye, L. Györfi, and G, Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York 1996.

[10] I. Dhillon, S. Mallela, and D. Modha. Information-Theoretic Co-clustering. *Proceedings of ACM SIGKDD'03*, 89 - 98, 2003

[11] S.Dudoit and J. Fridlyand. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9): 1090 - 1099, 2003.

[12] G.R. Grimmett and C.J.H. McDiarmid. On colouring random graphs. *Math. Proc. Cam. Phil. Soc.*, 77: 313 - 324, 1975.

[13] J. Han, J. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. *Proceedings of ACM SIGMOD'00*, 1 - 12, 2000.

[14] D. J. Hand, H. Mannila and P. Smyth. *Principles of Data Mining*. MIT Press 2001

[15] R. Karp. *Probabilistic Analysis of Algorithms*. Class Notes, UC-Berkeley 1988.

[16] M. Koyutürk, W. Szpankowski,and A. Grama. Biclustering Gene-Feature Matrices for Statistically Significant Dense Patterns. *IEEE Computer Society Bioinformatics Conference*, 480 - 483, Stanford, 2004.

[17] T. Lange, V. Roth, M. Braun, and J. Buhmann. Stability-Based Validation of Clustering Solution. *Neural Computation*, 16(6): 1299 - 1323, 2004.

[18] J. Liu, S. Paulsen, W. Wang, A. B. Nobel, and J. Prins. Mining Approximate Frequent Itemsets from Noisy Data. *Proceedings of ICDM'05*, 721 - 724, 2005.

[19] J. Liu, S. Paulsen, X. Sun, W. Wang, A.B. Nobel, and J. Prins. Mining approximate frequent itemsets in the presence of noise: algorithm and analysis. To appear in *Proceedings of SDM* 2006.

[20] D. Matula. The largest clique size in a random graph. Southern Methodist University, *Tech. Report*, CS 7608, 1976.

[21] S. Madeira and A. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1(1): 24 - 45, 2004

[22] M. Okamoto. Some inequalities relating to the partial sum of binomial probabilities. *Annals of the Institute of Statistical Mathematics*, 10: 29 - 35, 1958

[23] J. Pei, A.K. Tung, and J. Han. Fault-tolerant frequent pattern mining: Problems and challenges. *Proceedings of DMKD'01*, 2001.

[24] J. Pei, G. Dong, W. Zou, and J. Han. Mining Condensed Frequent-Pattern Bases. *Knowledge and Information Systems*, 6(5): 570 - 594, 2002.

[25] G. Park and W. Szpankowshi. Analysis of biclusters with applications to gene expression data. *Proceeding of AoA'05*, 2005.

[26] J. D. Reuning-Scherer. Mixture Models for Block Clustering. Phd Thesis, Yale university 1997.

[27] J. K. Seppänen, and H. Mannila. Dense Itemsets. *Proceedings of ACM SIGKDD'04*, 683 - 688, 2004.

[28] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(1): 136 - 144, 2002.

[29] A. Tanay, R. Sharan and R. Shamir. Biclustering Algorithms: A Survey. In *Handbook of Computational Molecular Biology*, Chapman & Hall/CRC, Computer and Information Science Series, 2005. In press.

[30] R. Tibshirani, G. Walther and T. Hastie. Estimating the number of clusters in a dataset via gap statistic. *Technical Report 208*, Dept of Statistics, Stanford University, 2000.

[31] C. Yang, U. Fayyad, and P. S. Bradley. Efficient discovery of error-tolerant frequent itemsets in high dimensions. *Proceedings of ACM SIGKDD'01*, 194 - 203, 2001.