Towards Average Case Analysis of Itemset Mining.

Dan Singer, David J. Haglin, and Anna M. Manning

***Abstract*— We perform a statistical analysis and describe the asymptotic behavior of the frequency and size distribution of δ-occurrent, minimal δ-occurrent, and maximal δ-occurrent itemsets occurring in random datasets across the entire spectrum of δ. We also describe the probability distribution of the support of an n-element itemset in a random dataset. Finally, we exhibit a class of datasets with an exponential number of minimal unique itemsets. We find that for small values of δ relative to number of transactions the size distribution of δ-occurrent itemsets and maximal δ-occurrent itemsets can be approximated by the binomial distributions $b(L, \frac{1}{1+2^\delta})$ and $b(L, \frac{1}{2^\delta})$, respectively, where L is inventory size. The ratio of minimal δ-occurrent and maximal δ-occurrent itemsets to the total number of δ-occurrent itemsets is low for small values of δ and rapidly approaches 1 as δ approaches the number of transactions. We also prove that the probability distribution of the support of an n-element itemset in a random k-transaction dataset is binomial of type $b(k, \frac{1}{2^n})$.**

Keywords: frequent itemsets, combinatorial properties, statistics, average case analysis

I. Introduction

Itemset mining is an important and well-studied branch of data mining. There have been hundreds of papers as well as workshops at conferences (e.g. FIMI'03 and FIMI'04 at IEEE ICDM'03 and IEEE ICDM'04) devoted to this subject. Almost all of the research has focused on *frequent* itemset mining, but some recent attention has been given to *infrequent* itemset mining [1]–[3].

Since frequent itemset mining has so many practical applications such as finding association rules (cf. [4]–[7]), much of the analysis has involved performing timing tests on standard datasets. It would be helpful to have a more theoretical framework from which to assess algorithms. There has been some research into complexity-theoretic issues of mining frequent itemsets [1], [8], [9]. But these studies have focused on issues such as #$\mathcal{P}$-Completeness which is difficult to use in practical analysis. It would be ideal to perform *average case* rather than *worst case* analysis to support the abundance of experimental analysis available in the literature.

A missing ingredient necessary to conduct average case analysis has been the understanding of the expected number of itemsets with specific properties (such as maximal frequent) in a dataset of a specific size drawn uniformly randomly from the collection of all binary matrices of that size. In this paper we present the asymptotic behavior of the frequency and size distributions of three important categories of itemsets: δ-occurrent, minimal δ-occurrent, and maximal δ-occurrent. We also describe the probability distribution of the support of an n-element itemset in a random k-transaction dataset. The rest of the paper is organized as follows. Section II gives our basic data mining terminology. Section III provides the statements of the main results. In Section IV we prove those main results. In Section V we exhibit an example of a class of datasets which has an exponential number of minimal 1-occurrent itemsets. In Section VI we give our conclusions.

Dan Singer is with the Department of Mathematics and Statistics, Minnesota State University, Mankato, MN 56001, USA (dan.singer@mnsu.edu), fax: +1 507-389-6376.

David J. Haglin is with the Department of Computer and Information Sciences, Minnesota State University, Mankato, MN 56001, USA (david.haglin@mnsu.edu), fax: +1 507-389-6376.

Anna M. Manning is with the School of Computer Science, University of Manchester, Oxford Rd., Manchester, M13 9PL, UK (anna@manchester.ac.uk), fax: +44 161-275-6204.

II. Data Mining Terminology

Let $\mathcal{I} = \{x_1, x_2, \ldots, x_L\}$ be an inventory of L *items*. An *itemset* is a subset $I \subseteq \mathcal{I}$. The *cardinality* of I, denoted by $|I|$, is the number of items in the itemset. A $k \times L$ *dataset* $\mathcal{D} = \{t_1, t_2, \ldots, t_k\}$ consists of a set of k *transactions* of the form $t_i = (i, T_i)$, where i is the *transaction identification* (tid) number of t_i and T_i is a subset of $\mathcal{I}$. We denote by $|\mathcal{D}|$ the number of transactions in the dataset. The *support set* of an itemset I with respect to the dataset $\mathcal{D}$ is

$$\mathcal{D}(I) = \{t_i \in \mathcal{D} : I \subseteq T_i\}.$$

The *support* of an itemset I in dataset $\mathcal{D}$ is the cardinality of the support set of I. That is, $Supp^{\mathcal{D}}(I) = |\mathcal{D}(I)|$. The *relative support* of an itemset, defined as $Supp^{\mathcal{D}}(I)/|\mathcal{D}|$, is a number between 0 and 1 inclusive.

The itemset I is said to be:

δ-*occurrent*	if	$Supp^{\mathcal{D}}(I) = \delta$
δ-*frequent*	if	$Supp^{\mathcal{D}}(I) \geq \delta$
δ-*infrequent*	if	$Supp^{\mathcal{D}}(I) < \delta$

In addition, an itemset is:

- *minimal* δ-*occurrent* if it is δ-occurrent and all of its proper subsets are $(\delta+1)$-frequent;
- *minimal* δ-*infrequent* if it is δ-infrequent and all of its proper subsets are $(\delta+1)$-frequent;
- *maximal* δ-*occurrent* if it is δ-occurrent and all of its proper supersets are δ-infrequent; and
- *maximal* δ-*frequent* if it is δ-frequent and all of its proper supersets are δ-infrequent.

Example: Let $\mathcal{I} = \{x_1, x_2, x_3, x_4, x_5\}$ and let $\mathcal{D} = \{t_1, t_2, t_3, t_4, t_5, t_6\}$ be the 6×5 dataset given by:

$$\begin{aligned} T_1 &= \{\}, \\ T_2 &= \{x_1\}, \\ T_3 &= \{x_1, x_2\}, \\ T_4 &= \{x_1, x_2, x_3\}, \\ T_5 &= \{x_1, x_2, x_3, x_4\}, \end{aligned}$$

$$T_6 = \{x_1, x_2, x_3\}.$$

The 3-occurrent itemsets are:

$$\begin{aligned} I_1 &= \{x_3\}, \\ I_2 &= \{x_1, x_3\}, \\ I_3 &= \{x_2, x_3\}, \\ I_4 &= \{x_1, x_2, x_3\}. \end{aligned}$$

All of these itemsets have support set $\{t_4, t_5, t_6\}$. Itemset I_1 is minimal 3-occurrent and itemset I_4 is maximal 3-occurrent.

We will represent $k \times L$ datasets by $k \times L$ binary matrices, where the entry in row i, column j is 1 if and only if transaction t_i contains item x_j. The example above can be represented by the matrix

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

We can see that there are 2^{kL} possible datasets of size k, drawn from an inventory of size L.

III. Statement of Main Results

In Theorem 1 we give asymptotic bounds for the expected number of δ-occurrent, minimal δ-occurrent, and maximal δ-occurrent itemsets in a random $k \times L$ dataset. In Theorem 2 we give asymptotic bounds for the expected size of a δ-occurrent, minimal δ-occurrent, and maximal δ-occurrent itemset in a random $k \times L$ dataset. We also provide asymptotic bounds for the standard deviation of the expected size of a δ-occurrent and maximal δ-occurrent itemset. In Theorem 3 we describe the probability distribution of the support of an n-element itemset in a random k-transaction dataset.

Theorem 1: Let $\mathcal{I}$ be an inventory of size L. Let $\Omega_{k,L}$ be the sample space consisting of $k \times L$ datasets of $\mathcal{I}$. Assume that the events in $\Omega_{k,L}$ are equally likely. For each $\delta \geq 1$ let $X_{k,L,\delta}$, $X^{min}_{k,L,\delta}$, and $X^{max}_{k,L,\delta}$ be the random variables on $\Omega_{k,L}$ defined respectively as the number of δ-occurrent, minimal δ-occurrent, and maximal δ-occurrent itemsets of a dataset $\mathcal{D} \in \Omega_{k,L}$.

(1) For any fixed k, L, and $\delta \geq 1$,

$$E[X_{k,L,\delta}] < \binom{k}{\delta}(1 + 2^{-\delta})^L$$

and

$$E[X^{max}_{k,L,\delta}] < \binom{k}{\delta}.$$

(2) For any fixed ratio $r > 0$ and fixed $\delta \geq 1$, numerical evidence suggests that

$$E[X_{rL,L,\delta}] = (1 - o(1))\binom{rL}{\delta}\left(1 + 2^{-\delta}\right)^L$$

and

$$E[X^{max}_{rL,L,\delta}] = (1 - o(1))\binom{rL}{\delta}$$

as $L \to \infty$. The rate of convergence decreases as r increases and as δ increases.

(3) For any fixed k, L, and $\delta \geq 1$,

$$e^{-\frac{L}{2^\delta - 1}} < \frac{E[X^{max}_{k,L,\delta}]}{E[X_{k,L,\delta}]} \leq \frac{E[X^{min}_{k,L,\delta}]}{E[X_{k,L,\delta}]} \leq 1.$$

Theorem 2: Let $\mathcal{I}$ be an inventory of size L. For each k and δ let $\Omega_{k,L,\delta}$ be the sample space consisting of ordered pairs of the form $(\mathcal{D}, I)$, where $\mathcal{D}$ is a $k \times L$ dataset of $\mathcal{I}$ and I is a δ-occurrent itemset of $\mathcal{D}$. Let $\Omega^{min}_{k,L,\delta}$ and $\Omega^{max}_{k,L,\delta}$ be the sample spaces consisting of dataset-itemset pairs in which the itemset is minimal and maximal δ-occurrent, respectively. Assume that the events in these sample spaces are equally likely. For each k, L, and δ let $Y_{k,L,\delta}$ be the random variable on $\Omega_{k,L,\delta}$ defined by

$$Y_{k,L,\delta}(\mathcal{D}, I) = |I|.$$

Let $Y^{min}_{k,L,\delta}$ and $Y^{max}_{k,L,\delta}$ be defined similarly on $\Omega^{min}_{k,L,\delta}$ and $\Omega^{max}_{k,L,\delta}$.

(1) For any fixed ratio $r > 0$ and fixed $\delta \geq 1$, numerical evidence suggests that the probability distribution of $Y_{rL,L,\delta}$ can be approximated by a binomial distribution of the form $b(L, \frac{1}{1+2^\delta})$. In particular, the mean is

$$E[Y_{rL,L,\delta}] = (1 + o(1))\frac{L}{1 + 2^\delta}$$

and the standard deviation is

$$\sigma_{rL,L,\delta} = (1 - o(1))\frac{\sqrt{L}}{2^{\delta/2} + 2^{-\delta/2}}$$

as $L \to \infty$. The rate of convergence for each expression decreases as r increases and as δ increases.

(2) For any fixed ratio $r > 0$ and fixed $\delta \geq 1$, numerical evidence suggests that the probability distribution of $Y^{max}_{rL,L,\delta}$ can be approximated by a binomial distribution of the form $b(L, \frac{1}{2^\delta})$. In particular, the mean is

$$E[Y^{max}_{rL,L,\delta}] = (1 + o(1))\frac{L}{2^\delta}$$

and the standard deviation is

$$\sigma^{max}_{rL,L,\delta} = (1 - o(1))\sqrt{L(2^{-\delta} - 4^{-\delta})}$$

as $L \to \infty$. The rate of convergence for each expression decreases as r increases and as δ increases.

(3) For any fixed k, L, and $\delta \geq 1$,

$$1 \leq \frac{E[Y^{max}_{k,L,\delta}]}{E[Y_{k,L,\delta}]} < e^{\frac{L}{2^\delta - 1}}.$$

Numerical evidence suggests that $E[Y^{min}_{k,L,\delta}]$ rapidly approaches $E[Y_{k,L,\delta}]$ as $\delta \to k$.

Theorem 3: Let $\mathcal{I}$ be an inventory of size L. For each k and $n \leq L$ let $\Omega^{(n)}_{k,L}$ be the sample space consisting of

ordered pairs of the form $(\mathcal{D}, I)$, where $\mathcal{D}$ is a $k \times L$ dataset of $\mathcal{I}$ and I is an itemset of $\mathcal{D}$ of cardinality n. Assume that the events in these sample spaces are equally likely. Let $Z_{k,L}^{(n)}$ be the random variable on $\Omega_{k,L}^{(n)}$ be defined by

$$Z_{k,L}^{(n)}(\mathcal{D}, I) = Supp^{\mathcal{D}}(I).$$

Then the probability distribution of $Z_{k,L}^{(n)}$ is binomial of the form $b(k, \frac{1}{2^n})$ with mean

$$E[Z_{k,L}^{(n)}] = \frac{k}{2^n}$$

and standard deviation

$$\sigma_{k,L}^{(n)} = \sqrt{k(2^{-n} - 4^{-n})}.$$

IV. Proofs

The key to proving Theorems 1, 2, and 3 is contained in Lemma 1, in which we count dataset-itemset pairs using standard methods of enumerative combinatorics: the rearrangements formula, generating functions, and the principle of inclusion-exclusion [10], [11]. We will prove Lemma 1 first, then use it to establish Theorems 1 through 3. In Theorems 2 and 3 we refer to the binomial distribution. For a reference, consult [12].

Lemma 1: Let $\mathcal{I}$ be an inventory of size L, and let I be a fixed itemset of size n drawn from $\mathcal{I}$. Let $f(k, L, \delta, n)$ denote the number of $k \times L$ datasets in which I is δ-occurrent. Then

$$f(k, L, \delta, n) = \binom{k}{\delta} 2^{(L-n)\delta} (2^L - 2^{L-n})^{k-\delta}.$$

If $f^{min}(k, L, \delta, n)$ and $f^{max}(k, L, \delta, n)$ denote the number of $k \times L$ datasets in which I is minimal and maximal δ-occurrent then

$$f^{min}(k, L, \delta, n) =$$

$$\binom{k}{\delta} 2^{(L-n)\delta} \sum_{j=0}^{n} (-1)^j \binom{n}{j} (2^L - (j+1)2^{L-n})^{k-\delta}$$

and

$$f^{max}(k, L, \delta, n) = (1 - 2^{-\delta})^{L-n} f(k, L, \delta, n).$$

A. Proof of Lemma 1

Assume that $I = \{x_{s_1}, x_{s_2}, \ldots, x_{s_n}\}$. In order to count all datasets $\mathcal{D}$ in which I is δ-occurrent, we must count all $k \times L$ binary matrices in which exactly δ rows have 1s in columns s_1, s_2, ..., s_n. There are 2^L types of row which can appear in any binary matrix with L columns, corresponding to the number of binary strings of length L. There are 2^{L-n} row types which contain 1s in positions s_1 through s_n, and we will label these P_1, P_2, ..., $P_{2^{L-n}}$. There are $2^L - 2^{L-n}$ other row types, which we will label Q_1, Q_2, ..., $Q_{2^L - 2^{L-n}}$. To construct all possible binary matrices containing exactly δ rows with 1s in columns s_1 through s_n, we will first choose p_i copies of P_i for each $i \leq 2^{L-n}$, then q_i copies of Q_i for each $i \leq 2^L - 2^{L-n}$, requiring that $\sum_{i=1}^{2^{L-n}} p_i = \delta$ and $\sum_{i=1}^{2^L - 2^{L-n}} q_i = k - \delta$, then choose all possible rearrangements of these row types within a $k \times L$ matrix, then sum over all possibilities. Using the rearrangements formula, the total number is

$$\sum_{\sum p_i = \delta} \sum_{\sum q_i = k-\delta} \frac{k!}{p_1! \cdots p_{2^{L-n}}! q_1! \cdots q_{2^L - 2^{L-n}}!},$$

which we recognize as $k!$ times the coefficient of x^δ in the generating function $(e^x)^{2^{L-n}}$ times the coefficient of $x^{k-\delta}$ in the generating function $(e^x)^{2^L - 2^{L-n}}$, namely

$$f(k, L, \delta, n) = k! \frac{(2^{L-n})^\delta}{\delta!} \frac{(2^L - 2^{L-n})^{k-\delta}}{(k-\delta)!}.$$

To compute $f^{min}(k, L, \delta, n)$, we must subtract from $f(k, L, \delta, n)$ the total number of datasets $\epsilon^{min}(k, \delta, n)$ in which I contains a subset J of size $n-1$ which is also δ-occurrent in $\mathcal{D}$. This requires an inclusion-exclusion argument. If we denote by A_i the set of all datasets in which I and $I - \{x_{s_i}\}$ are δ-occurrent, then

$$\epsilon^{min}(k, \delta, n) = |A_1 \cup \cdots \cup A_n|.$$

The datasets in $A_{e_1} \cap \cdots \cap A_{e_j}$ are those in which the itemsets I, $I - \{x_{s_{e_1}}\}$, ..., $I - \{x_{s_{e_j}}\}$ are all δ-occurrent. As above there are 2^{L-n} row types P_1 through $P_{2^{L-n}}$ which contain a 1 in positions s_1 through s_n. For each $i \leq j$ there are 2^{L-n} row types which have 1s in all positions of $\{s_1, \ldots, s_n\}$ except s_{e_i} and a 0 in position s_{e_i}, and we must not choose any of these row types. There are $2^L - (j+1)2^{L-n}$ remaining row types, which we will label R_1 through $R_{2^L - (j+1)2^{L-n}}$. To construct all possible binary matrices representing datasets in $A_{e_1} \cap \cdots \cap A_{e_j}$, we will first choose p_i copies of P_i for each $i \leq 2^{L-n}$, then q_i copies of R_i for each $i \leq 2^L - (j+1)2^{L-n}$, requiring that $\sum_{i=1}^{2^{L-n}} p_i = \delta$ and $\sum_{i=1}^{2^L - (j+1)2^{L-n}} q_i = k - \delta$, then choose all possible rearrangements of these row types within a $k \times L$ matrix, then sum over all possibilities. The total number is

$$|A_{e_1} \cap \cdots \cap A_{e_j}| = k! \frac{(2^{L-n})^\delta}{\delta!} \frac{(2^L - (j+1)2^{L-n})^{k-\delta}}{(k-\delta)!}.$$

This, combined with the inclusion-exclusion formula, yields $f^{min}(k, L, \delta, n)$.

To compute $f^{max}(k, L, \delta, n)$, we must subtract from $f(k, L, \delta, n)$ the total number of datasets $\epsilon^{max}(k, L, \delta, n)$ in which I is contained in a superset J of size $n+1$ which is also δ-occurrent in $\mathcal{D}$. This requires another inclusion-exclusion argument. Write $I^c = \{x_{t_1}, \ldots, x_{t_{L-n}}\}$. If we denote by B_i the set of all datasets in which I and $I \cup \{x_{t_i}\}$ are δ-occurrent, then

$$\epsilon^{max}(k, L, \delta, n) = |B_1 \cup \cdots \cup B_{L-n}|.$$

The datasets in $B_{e_1} \cap \cdots \cap B_{e_j}$ are those in which the itemsets I, $I \cup \{x_{t_{e_1}}\}$, ..., $I \cup \{x_{t_{e_j}}\}$ are all δ-occurrent, i.e. in which $I \cup \{x_{t_{e_1}}, \ldots, x_{t_{e_j}}\}$ is δ-occurrent. There are 2^{L-n-j} row types with 1s in the positions of $\{s_1, \ldots, s_n, t_{e_1}, \ldots, t_{e_j}\}$, and we must choose δ rows from among these types. There are $2^L - 2^{L-n}$ row types which do not have 1s in all

the positions of $\{s_1, \ldots, s_n\}$, and we must choose $k - \delta$ rows from among these types. Summing over all possible rearrangements as before we obtain

$$|B_{e_1} \cap \cdots \cap B_{e_j}| = k! \frac{(2^{L-n-j})^\delta}{\delta!} \frac{(2^L - 2^{L-n})^{k-\delta}}{(k-\delta)!}.$$

This, combined with the inclusion-exclusion formula, yields

$$\epsilon^{max}(k, L, \delta, n) =$$

$$\binom{k}{\delta} \sum_{j=1}^{L-n} (-1)^{j-1} \binom{L-n}{j} (2^{L-n-j})^\delta (2^L - 2^{L-n})^{k-\delta}.$$

Therefore

$$f^{max}(k, L, \delta, n) =$$

$$\binom{k}{\delta} \sum_{j=0}^{L-n} (-1)^{j} \binom{L-n}{j} (2^{L-n-j})^\delta (2^L - 2^{L-n})^{k-\delta} =$$

$$\binom{k}{\delta} 2^{(L-n)\delta} (2^L - 2^{L-n})^{k-\delta} \sum_{j=0}^{L-n} (-1)^{j} \binom{L-n}{j} (2^{-\delta})^j =$$

$$f(k, L, \delta, n)(1 - 2^{-\delta})^{L-n}.$$

B. Proof of Theorem 1

$E[X_{k,L,\delta}]$ is the ratio of the number of dataset-itemset pairs $(\mathcal{D}, I)$ in which I is δ-occurrent in $\mathcal{D}$ to the size of the sample space, hence by Lemma 1 we have

$$E[X_{k,L,\delta}] = |\Omega_{k,L}|^{-1} \sum_{n=0}^{L} \binom{L}{n} f(k, L, \delta, n) =$$

$$2^{-kL} \binom{k}{\delta} \sum_{n=0}^{L} \binom{L}{n} 2^{(L-n)\delta} (2^L - 2^{L-n})^{k-\delta} =$$

$$\binom{k}{\delta} \sum_{n=0}^{L} \binom{L}{n} 2^{-n\delta} (1 - 2^{-n})^{k-\delta} < \binom{k}{\delta} \sum_{n=0}^{L} \binom{L}{n} 2^{-n\delta} =$$

$$\binom{k}{\delta} (1 + 2^{-\delta})^L.$$

Similarly, using

$$f^{max}(k, L, \delta, n) = (1 - 2^{-\delta})^{L-n} f(k, L, \delta, n),$$

we have

$$E[X^{max}_{k,L,\delta}] < \binom{k}{\delta} \sum_{n=0}^{L} \binom{L}{n} (1 - 2^{-\delta})^{L-n} 2^{-n\delta} = \binom{k}{\delta}.$$

A plot of $y = E[X_{rL,L,\delta}] / \binom{rL}{\delta} (1 + 2^{-\delta})^L$ versus L for various choices of r and δ suggests that $y \to 1$ as $L \to \infty$. The rate of convergence decreases as r increases and as δ increases. We illustrate this for $r = .25, .5, 1, 2, 4$ and $\delta = 1$ in Figure 1.

A plot of $y = E[X^{max}_{rL,L,\delta}] / \binom{rL}{\delta}$ versus L for various choices of r and δ suggests that $y \to 1$ as $L \to \infty$. The rate of convergence decreases as r increases and as δ increases. We illustrate this for $r = .25, .5, 1, 2, 4$ and $\delta = 2$ in Figure 2.

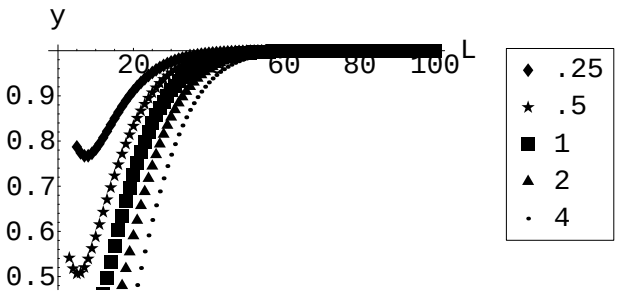

Fig. 1. $y = E[X_{rL,L,1}] / rL(1.5)^L$ for $r = .25, .5, 1, 2, 4$

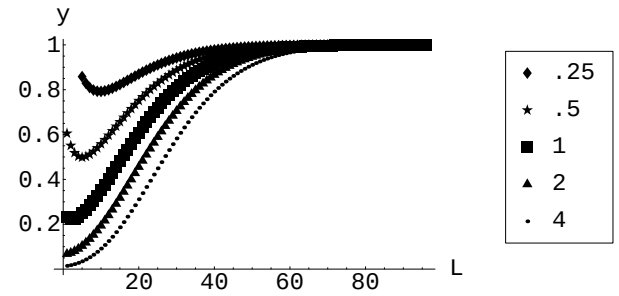

Fig. 2. $y = E[X^{max}_{rL,L,2}] \Big/ \binom{rL}{2}$ for $r = .25, .5, 1, 2, 4$

Observe that every δ-occurrent itemset of a dataset is contained in exactly one maximal δ-occurrent itemset. The reason for this is that if the support set of a δ-occurrent itemset I is $\{t_{e_1}, \ldots, t_{e_\delta}\}$, then the maximal δ-occurrent itemset J which has this support set is uniquely determined by the equation

$$J = \bigcap_{i=1}^{\delta} T_{e_i}.$$

This implies

$$\sum_{n=0}^{L} \binom{L}{n} f^{max}(k, L, \delta, n) \leq \sum_{n=0}^{L} \binom{L}{n} f^{min}(k, L, \delta, n)$$

for every k, L, and δ. This, combined with

$$f^{max}(k, L, \delta, n) = (1 - 2^{-\delta})^{L-n} f(k, L, \delta, n)$$

and

$$(1 - 2^{-\delta})^{L-n} \geq (1 - 2^{-\delta})^L > e^{-\frac{L}{2^\delta - 1}},$$

proves part (3) of Theorem 1.

C. Proof of Theorem 2

By Lemma 1, there are $\binom{L}{n} f(k, L, \delta, n)$ dataset-itemset pairs $(\mathcal{D}, I)$ in which $|I| = n$ and I is δ-occurrent in $\mathcal{D}$. Therefore

$$|\Omega_{k,L,\delta}| = \sum_{n=0}^{L} \binom{L}{n} f(k, L, \delta, n)$$

and

$$P[Y_{k,L,\delta} = n] = |\Omega_{k,L,\delta}|^{-1} \binom{L}{n} f(k, L, \delta, n).$$

We can compute $E[Y_{k,L,\delta}]$ using these expressions. For example, $E[Y_{1000,100,1}] = 33.3335$ and $E[Y_{1000,100,2}] = 20.1366$. See Figures 3 and 4 for a graph of $y = E[Y_{1000,100,\delta}]$ versus δ for $3 \le \delta \le 100$.

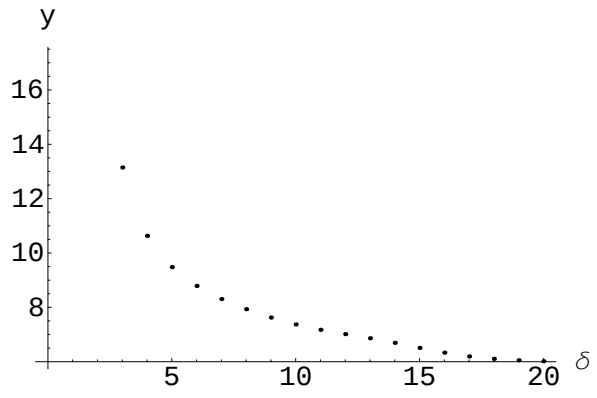

Fig. 3. $y = E[Y_{1000,100,\delta}]$, $3 \le \delta \le 20$

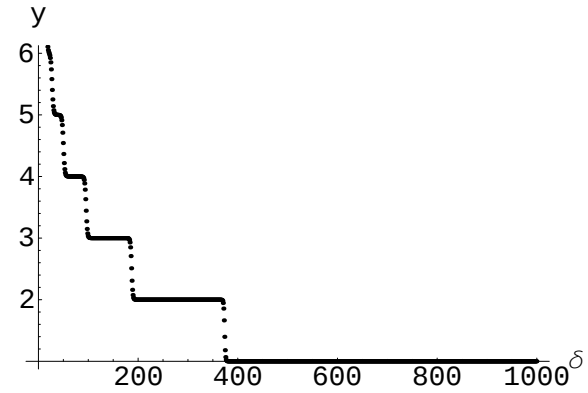

Fig. 4. $y = E[Y_{1000,100,\delta}]$, $20 \le \delta \le 1000$

By Theorem 1 we can say that

$$|\Omega_{rL,L,\delta}| = (1 - o(1))2^{rL^2}\binom{rL}{\delta}(1 + 2^{-\delta})^L.$$

Therefore

$$P[Y_{rL,L,\delta} = n] =$$
$$(1 + o(1))\frac{\binom{L}{n}\binom{rL}{\delta}2^{(L-n)\delta}(2^L - 2^{L-n})^{rL-\delta}}{2^{rL^2}\binom{rL}{\delta}(1 + 2^{-\delta})^L}.$$

Simplifying, this can be expressed in the form

$$P[Y_{rL,L,\delta} = n] = (1+o(1))(1-2^{-n})^{rL-\delta}\binom{L}{n}p^n(1-p)^{L-n}$$

where $p = \frac{1}{1+2^\delta}$. This can be approximated by $b(L, p)$, the binomial distribution, by dropping the

$$(1 + o(1))(1 - 2^{-n})^{rL-\delta}$$

term. This yields approximations to the mean and standard deviation of $Y_{rL,L,\delta}$ for small values of δ relative to number of transactions:

$$E[Y_{rL,L,\delta}] \approx \frac{L}{1 + 2^\delta}$$

and

$$\sigma_{rL,L,\delta} \approx \frac{\sqrt{L}}{2^{\delta/2} + 2^{-\delta/2}}.$$

To check the accuracy of these approximations we plotted

$$y = E[Y_{rL,L,\delta}] \Big/ \frac{L}{1 + 2^\delta}$$

and

$$y = \sigma_{rL,L,\delta} \Big/ \frac{\sqrt{L}}{2^{\delta/2} + 2^{-\delta/2}}$$

versus L for various values of r and δ and observed that in all cases $y \to 1$ as $L \to \infty$. The rate of convergence decreases as r increases and as δ increases. See Figures 5 and 6 for $r = .25, .5, 1, 2, 4$ and $\delta = 1$.

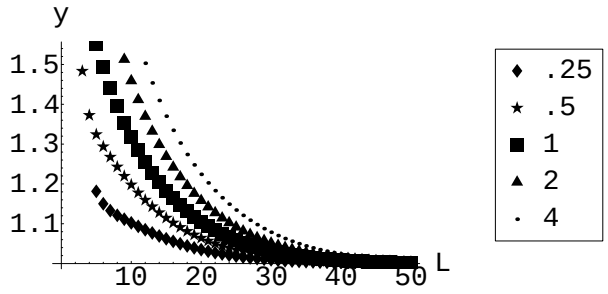

Fig. 5. $y = E[Y_{rL,L,1}] \Big/ \frac{L}{3}$ for $r = .25, .5, 1, 2, 4$

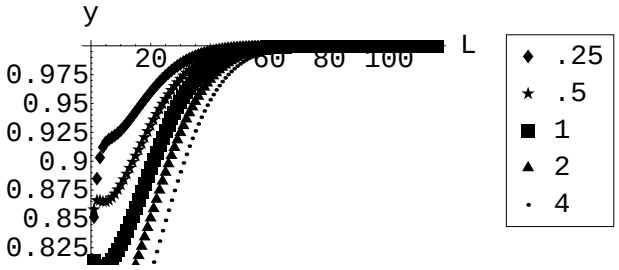

Fig. 6. $y = \sigma_{rL,L,1} \Big/ \frac{\sqrt{L}}{2^{1/2}+2^{-1/2}}$ for $r = .25, .5, 1, 2, 4$

Similarly, we have

$$P[Y^{max}_{rL,L,\delta} = n] =$$
$$(1 + o(1))\frac{\binom{L}{n}\binom{rL}{\delta}2^{(L-n)\delta}(2^L - 2^{L-n})^{rL-\delta}(1 - 2^{-\delta})^{L-n}}{2^{rL^2}\binom{rL}{\delta}}.$$

Simplifying, this can be expressed in the form

$$P[Y_{rL,L,\delta} = n] = (1+o(1))(1-2^{-n})^{rL-\delta}\binom{L}{n}p^n(1-p)^{L-n}$$

where $p = \frac{1}{2^\delta}$. This can be approximated by $b(L, p)$, the binomial distribution. This yields approximations to the mean

and standard deviation of $Y_{rL,L,\delta}$ for small values of δ relative to number of transactions:

$$E[Y_{rL,L,\delta}] \approx \frac{L}{2^\delta}$$

and

$$\sigma^{max}_{rL,L,\delta} \approx \sqrt{L(2^{-\delta} - 4^{-\delta})}.$$

To check the accuracy of these approximations we plotted

$$y = E[Y^{max}_{rL,L,\delta}] \Big/ \frac{L}{2^\delta}$$

and

$$y = \sigma^{max}_{rL,L,\delta} \Big/ \sqrt{L(2^{-\delta} - 4^{-\delta})}$$

versus L for various values of r and δ and observed that in all cases $y \to 1$ as $L \to \infty$. The rate of convergence decreases as r increases and as δ increases. See Figures 7 and 8 for $r = .25, .5, 1, 2, 4$ and $\delta = 2$.

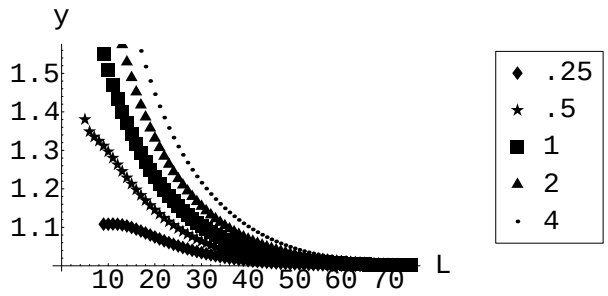

Fig. 7. $y = E[Y^{max}_{rL,L,2}] \Big/ \frac{L}{4}$ for $r = .25, .5, 1, 2, 4$

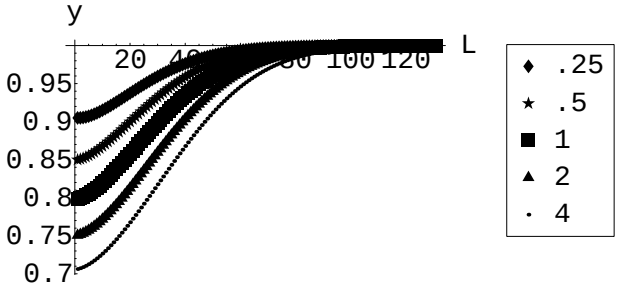

Fig. 8. $y = \sigma^{max}_{rL,L,2} \Big/ \sqrt{L(2^{-2} - 4^{-2})}$ for $r = .25, .5, 1, 2, 4$

To prove the inequality in part (3) of Theorem 2, we combine

$$E[Y^{max}_{k,L,\delta}] = \frac{\sum_{n=0}^{L} n\binom{L}{n} f^{max}(k,L,\delta,n)}{\sum_{n=0}^{L}\binom{L}{n} f^{max}(k,L,\delta,n)} <$$

$$\frac{\sum_{n=0}^{L} n\binom{L}{n} f(k,L,\delta,n)}{(1-2^{-\delta})^L \sum_{n=0}^{L}\binom{L}{n} f(k,L,\delta,n)} = (1-2^{-\delta})^{-L} E[Y_{k,L,\delta}]$$

with

$$(1-2^{-\delta})^{-L} < e^{\frac{L}{2^{\delta-1}}}.$$

Numerical evidence suggests that both $E[Y^{max}_{k,L,\delta}]$ and $E[Y^{min}_{k,L,\delta}]$ rapidly approach $E[Y_{k,L,\delta}]$ as $\delta \to k$. See Figure 9 for an illustration of this when $k = 1000$ and $L = 100$, in which we superimposed the graphs of

$$y = E[Y^{min}_{1000,100,\delta}]/E[Y_{1000,100,\delta}]$$

and

$$y = E[Y^{max}_{1000,100,\delta}]/E[Y_{1000,100,\delta}]$$

for $1 \le \delta \le 7$.

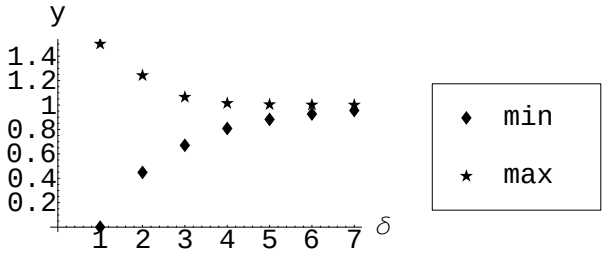

Fig. 9. Expected min, max itemset size ratios, $k = 1000$, $L = 100$

D. Proof of Theorem 3

This follows from

$$|\Omega^{(n)}_{k,L}| = \binom{L}{n} 2^{kL}$$

and

$$P[Z^{(n)}_{k,L} = \delta] = \frac{\binom{L}{n} f(k,L,\delta,n)}{|\Omega^{(n)}_{k,L}|} = \binom{k}{\delta} 2^{-n\delta}(1-2^{-n})^{k-\delta}.$$

V. Example of a Class of Datasets with an Exponential Number of Minimal Unique Itemsets

For each positive odd integer $L \ge 3$ let $\mathcal{I}_L = \{x_0, \ldots, x_{L-1}\}$ be an itemset of size L and let $\mathcal{D}_L$ denote the $L \times L$ dataset constructed as follows:

$$\mathcal{D}_L = \{t_0, t_1, \ldots, t_{L-1}\}$$

where

$$T_{2i} = T_{2i+1} = \{x_0, \ldots, x_{L-1}\} - \{x_{2i}, x_{2i+1}\}$$

for $0 \le i \le \frac{L-3}{2}$ and

$$T_{L-1} = \{x_0, \ldots, x_{L-1}\}.$$

For example, the binary matrix which represents $\mathcal{D}_5$ is

$$\begin{bmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

The transaction t_{L-1} is contained in the support set of every itemset. Therefore the 1-occurrent itemsets are those that contain at least one item from $\{x_{2i}, x_{2i+1}\}$ for each

$0 \leq i \leq \frac{L-3}{2}$ and an arbitrary subset of $\{x_{L-1}\}$. There are $2 \times 3^{\frac{L-1}{2}} = \frac{2}{\sqrt{3}}(\sqrt{3})^L$ 1-occurrent itemsets, exactly $2^{\frac{L-1}{2}} = \frac{1}{\sqrt{2}}(\sqrt{2})^L$ of which are minimal 1-occurrent and exactly 1 of which is maximal 1-occurrent. The average size of a 1-occurrent itemset is

$$\left(2 \times 3^{\frac{L-1}{2}}\right)^{-1} \sum_{j=0}^{\frac{L-1}{2}} \binom{\frac{L-1}{2}}{j} 2^j (2L-1-2j) = \frac{2}{3}L - \frac{1}{6},$$

all the minimal 1-occurrent itemsets have size $\frac{1}{2}L - \frac{1}{2}$, and the unique maximal 1-occurrent itemset has size L.

VI. Conclusions

A. The expected number and size of maximal and minimal δ-occurrent itemsets when δ is small

For small values of δ relative to k there are far fewer minimal δ-occurrent itemsets and far fewer maximal δ-occurrent itemsets than there are δ-occurrent itemsets per random $k \times L$ dataset, and the expected size of maximal δ-occurrent itemsets is larger than the expected size of all δ-occurrent itemsets by a factor of roughly $1 + 2^{-\delta}$. As δ approaches k, the expected number and size of minimal and maximal δ-occurrent itemsets rapidly approaches the expected number and size of all δ-occurrent itemsets.

B. The variance of the expected size of δ-occurrent and maximal δ-occurrent itemsets for small values of δ

Since the probability distributions of the expected size of a δ-occurrent itemset and a maximal δ-occurrent itemset in a random dataset are approximately binomial for small values of δ relative to number of transactions, we can say that roughly 95% of all δ-occurrent itemsets will be of size within two standard deviations of the mean for small values of δ. For example, roughly 95% of all 1-occurrent itemsets occurring in random $k \times 900$ datasets will have size between 271 and 329, and roughly 95% of all maximal 1-occurrent itemsets will have size between 420 and 480.

C. Most δ-occurrent itemsets are the same size for large δ

For relatively large values of δ, the size distribution of a δ-occurrent itemset is very tightly clustered about the mean. For example, the expected size of a 200-occurrent itemset in a 1000×100 dataset is 2.000007, and the probability that a 200-occurrent itemset in a random 1000×100 dataset has size equal to 2 is 0.999993. So with a high degree of certainty we should say that almost all 200-occurrent itemsets in a random 1000×100 dataset have size equal to 2.

D. Statistics on δ-frequent itemsets

Statistics on δ-occurrent itemsets yield statistics on δ-frequent itemsets. For example, the probability that an itemset of size 10 is 7-frequent in a random 10000×100 dataset is found as follows: the mean support of a random itemset of size 10 is $\mu = 9.76563$ and the standard deviation is $\sigma = 3.12347$, therefore the z-score corresponding to 7 is

$$z = \frac{7 - \mu}{\sigma} = -0.885432.$$

Using a normal distribution with mean μ and standard deviation σ to approximate $b(10000, \frac{1}{2^{10}})$, we have

$$P[Z^{(10)}_{10000,100} \geq 7] \approx P[z \geq -0.885432] = 0.82.$$

E. For smaller δ there is higher variance of sizes and expected numbers of itemsets

The expected number and size distribution of δ occurrent itemsets in any given dataset may be far from what is predicted for a random dataset. As δ decreases, the variance increases. Consider the $L \times L$ dataset $\mathcal{D}_L$ we constructed in Section V above. There are $\frac{2}{\sqrt{3}}(\sqrt{3})^L$ 1-occurrent itemsets in $\mathcal{D}_L$, which is exponentially larger than the expected value of $(1 - o(1))L(1.5)^L$. There is 1 maximal 1-occurrent itemset, compared with an expected value of $(1-o(1))L$. The average size of a 1-occurrent itemset in $\mathcal{D}_L$ is $\frac{2}{3}L - \frac{1}{6}$, compared with the expected value of $(1 + o(1))\frac{1}{3}L$. The unique maximal 1-occurrent itemset has size L, compared with an expected value of $(1 + o(1))\frac{1}{2}L$.

References

[1] D. Gunopulos, R. Khardon, H. Mannila, S. Saluja, H. Toivonen, and R. S. Sharma, "Discovering all most specific sentences," *ACM Trans. Database Syst.*, vol. 28, no. 2, pp. 140–174, 2003.

[2] A. M. Manning and D. J. Haglin, "A new algorithm for finding minimal sample uniques for use in statistical disclosure assessment," in *IEEE International Conference on Data Mining (ICDM05)*, Nov. 2005, pp. 290–297.

[3] A. M. Manning, D. J. Haglin, and J. A. Keane, "A recursive search algorithm for statistical disclosure assessment," *Data Mining and Knowledge Discovery*, 2007, conditionally accepted.

[4] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 International Conference on Management of Data (SIGMOD 93)*, May 1993, pp. 207–216.

[5] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo, "Fast discovery of association rules," in *Advances in Knowledge Discovery and Data Mining, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Eds. The AAAI Press, Menlo Park*, 1996, pp. 307–328.

[6] S. Brin, R. Motwani, J. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," in *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*. ACM Press New York, NY, USA, 1997, pp. 255–264.

[7] M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New algorithms for fast discovery of association rules." in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1997, pp. 283–286.

[8] E. Boros, V. Gurvich, L. Khachiyan, and K. Makino, "On the complexity of generating maximal frequent and minimal infrequent sets," in *Symposium on Theoretical Aspects of Computer Science*, 2002, pp. 133–141. [Online]. Available: citeseer.ist.psu.edu/boros02complexity.html

[9] G. Yang, "Computational aspects of mining maximal frequent patterns," *Theoretical Computer Science*, vol. 362, pp. 63–85, 2006.

[10] R. P. Stanley, *Enumerative combinatorics. Vol. 1.* Cambridge: Cambridge University Press, 1997, with a foreword by Gian-Carlo Rota, Corrected reprint of the 1986 original.

[11] ——, *Enumerative combinatorics. Vol. 2.* Cambridge: Cambridge University Press, 1999, with a foreword by Gian-Carlo Rota and appendix 1 by Sergey Fomin.

[12] R. V. Hogg and E. A. Tanis, *Probability and Statistical Inference*. Prentice Hall, ISBN 0-13-146413-2, 2006.