Realistic synthetic data for testing association rule mining algorithms for market basket databases

ABSTRACT

Association rule mining (ARM) is an important subtask in Knowledge Discovery in Databases. Existing ARM algorithms have largely been tested using artificial data generated by the QUEST program developed by Agrawal et al. [2]. Concerns have been raised before [7, 25] on the significance of such sample data.

We provide the first theoretical investigation of the statistical properties of the databases generated by the QUEST program. Motivated by the claim (supported by empirical evidence) that item occurrences in real life market basket databases follow a rather different pattern, we then propose an alternative model for generating artificial data. We claim that such a model is simpler than QUEST and generates structures that are closer to real-life market basket data.

1. INTRODUCTION

The ARM problem is a well established topic in KDD. The problem addressed by ARM is to identify a set of relations (associations) in a binary valued attribute set which describe the likely coexistence of groups of attributes. To this end it is first necessary to identify *frequent* itemsets; those subsets F of the available set of attributes $\mathcal{I}$ for which the *support*, the number of times F occurs in the dataset under consideration, exceeds some threshold value. Other criteria are then applied to these itemsets to generate Association Rules (ARs) of the form $A \rightarrow B$, where A and B represent disjoint subsets of a frequent itemset F such that $A \cup B = F$.

A vast array of algorithms and techniques have been developed to solve this problem ([1, 2, 4, 7, 12, 19, 21] to cite but a few of the best known). However several fundamental issues are still unsolved in ARM. In particular the evaluation and comparison of ARM algorithms is a very difficult task [27], and it is often tackled by resorting to experiments carried out using data generated by the well established QUEST program from the IBM Quest Research Group [2]. The intricacy of this program makes it difficult to draw theoretical predictions on the behaviour of the various algorithms on input produced by this program. Empirical comparisons made in this way are also difficult to generalise because of the wide range of possible variation, both in the characteristics of the data (the structural characteristics of the synthetic data bases generated by QUEST are governed by a several interacting parameters), and in the environment in which the algorithms are being applied. It has also been noted [7] that data sets produced using the QUEST generator might be inherently not the hardest to deal with. In fact there is evidence that suggests that the performances of some algorithms on real data is much worse than those found on synthetic data generated using QUEST [25].

This paper presents three main results. First of all we provide additional arguments supporting the view that real-life databases show structural properties that are very different from those of the data generated by QUEST. There has been recent growing interest in various areas of Computer Science [3, 10, 20] in the class of so-called *heavy tail* statistical distributions. Distributions of this kind have been used in the past to describe word frequencies in text [26], the distribution of animal species [23] and of income [16], scientific citations count [18] and many other phenomena. We claim that it is at least reasonable to conjecture that similar distributions arise naturally also in the context of market basket databases. To support our claim we analyse empirically the distribution of item occurrences in four real-world retail databases publicly available on the web. In each case we suggest that the empirical distribution may fit (over a wide range of values) a heavy tailed distribution. Furthermore we argue that the data generated by QUEST shows quite different properties When the same empirical analysis mentioned above is performed on data generated by QUEST the results are quite different from those obtained for real-life retail databases. Differences have been found before [25] in the transaction sizes of the real-life vs QUEST generated databases. However some of these differences may be ironed out by a careful choice of the numerous parameters that controls the output of the QUEST generator. We contend that our analysis points to possible differences at a much deeper level.

Motivated by the outcomes of our empirical study, the second contribution of this paper is a mathematical study of the distribution of item occurrences in a typical large QUEST database. The study, at least in a simplified setting, confirms the empirical findings. The item occurrence distri-

bution in the databases generated by a simplified version of QUEST seems to follow a geometric (light tailed) rather than a sub-exponential or even polynomial (heavy tailed) decay. Other authors have studied the theoretical performances of some ARM algorithms [24] even assuming that the input databases are generated by some random process [17]. However, to the best of our knowledge, a study of the structural properties of the databases generated by QUEST (properties which may well be responsible for the observed [7, 25] particular behaviour of various mining algorithms on such datasets) has never been done before.

The final contribution of this paper is a proposal for an alternative synthetic data generator. Given the statistical features which seem to emerge from the real life databases it is natural to try and generate synthetic data that match such properties. Our model is reminiscent of the proposal put forward to model author citation networks and the web [3]. The main mechanism that leads to the desired properties is the so called *preferential attachment*, whereby successive transactions are filled by selecting items based on some given measure of their popularity (rather than at random). We complete our analysis by proving that the resulting databases show an asymptotic heavy tailed item occurrence distribution.

The rest of the paper has the following structure. Section 2 reports our empirical analysis of a number of real and synthetic databases. Section 3 presents the mathematical investigation of the structural properties of the databases generated by QUEST. Finally in Section 4 we describe our proposal for an alternative synthetic data generator.

2. REAL DATA ANALYSIS

From now on a database $\mathcal{D}$ will be a collection of h transactions, each being a set containing a certain number of items out of a collection $\mathcal{I}$ of n items. For $r \in \{0, \ldots, h\}$ let N_r be the number of items that occur in r transactions. In this section we substantiate the claim that, at least for market basket data, the sequence $(N_r)_{r \in \{0,\ldots,h\}}$ follows a distribution that has a rather "fat" tail and, on the contrary, the typical QUEST data shows rather different patterns. To this end we use the following data sets available from `http://fimi.cs.helsinki.fi/data/`.

`BMS-POS` First referred to in [25]. The database contains $515,597$ transactions on $1,657$ items. The average transaction size is 6.5 (max 164). The data was obtained from a large electronics retailer.

`BMS-WebView-1` First referred to in [25]. The database contains $59,602$ transactions on 497 items with average transaction size 2.5 (max 267). The dataset contains records of click-stream data from an e-commerce website. Each transaction represents a web session by all product detail pages viewed in that session.

`BMS-WebView-2` Similar scenario to the previous dataset, but the file contains $77,512$ transactions on $3,340$ items with average transaction size of 5 (max 161).

`retail.dat` First mentioned in [6]. The file contains $88,162$ transactions on $16,470$ items with average transaction

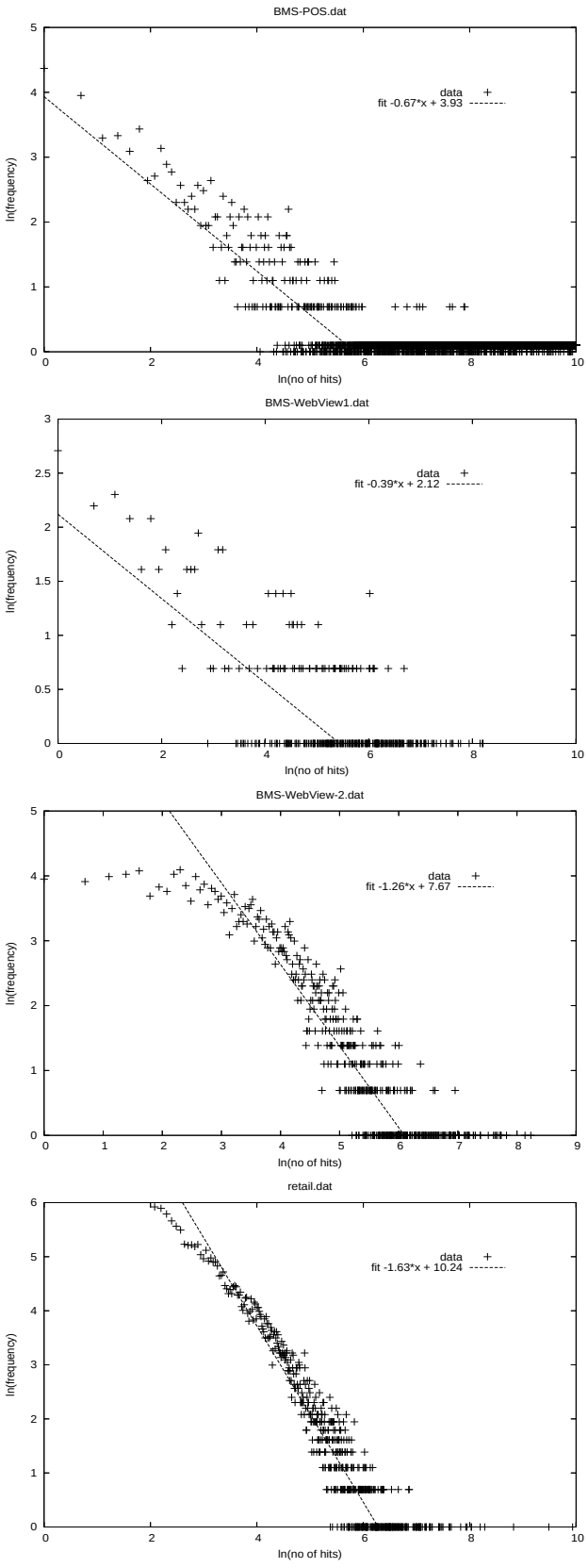

Figure 1: Log-log plots of the real-file data sets along with the best fitting lines

size 13. The data was obtained from a Belgian retail supermarket store.

`T10I4D100K.dat` This is a database generated with the program of Agrawal and Srikant. It contains 100,000 transactions on 1,000 items, with average transaction size 10.

`T40I10D100K.dat` Denser synthetic database. It still contains 100,000 transactions on 1,000 items, but the average transaction size is 40.

The plots in Figure 1 show (on a doubly logarithmic scale) the sequence $(N_r)_{r\in\{0,\ldots,h\}}$ in each case, along with the best fitting lines (computed through least square approximations using the `fit` command of `gnuplot`) over the whole range of values (approximate slope values are reported in each picture). Figure 2 shows the same statistics obtained using the two synthetic databases.

Given that all tested data-sets have homogeneous features (few hundreds to few thousands attributes, between around 50K and 500K transactions, small average transactions sizes and density) the differences between the plots in Figure 1 and 2 are striking. Although it may be argued that the number of real datasets examined is too small and the test carried out too coarse, our numerical calculations show that the sequences $(N_r)_{r\in\{0,\ldots,h\}}$ obtained from real-life market basket databases fit a straight line much better than the same sequences obtained from the synthetic QUEST databases. This leads us to conjecture that the two types of databases have in fact rather different properties. Furthermore the good line fitting of the real database sequences (especially in the tail of the sequences) leads to the additional conjecture that the studied distributions may be heavy tailed. More specifically, a distribution can be called *heavy tailed* if it decays with a sub-exponential rate [22]. For instance *power law* distributions decay like x^{-z} for some fixed $z > 0$. On a doubly logarithmic scale the relationship between the coordinates of a dataset having such decay would appear to be linear. This is exactly what happens in the case of the four market-basket datasets described above.

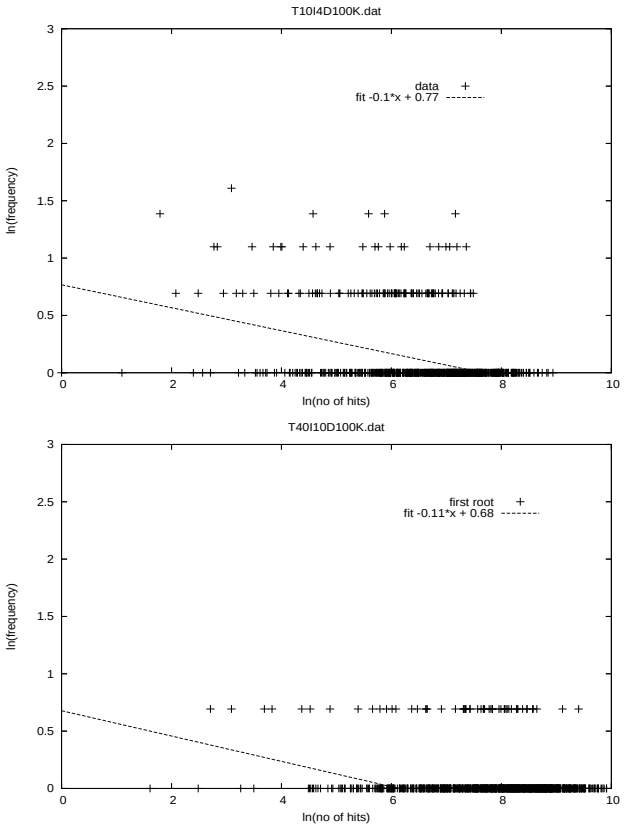

Figure 2: Log-log plots of the QUEST data sets along with the best fitting lines

3. A CLOSER LOOK AT QUEST

In this section we make quantitative estimates on some structural parameters related to the databases generated by the QUEST program. In particular we will provide further evidence supporting the claim that the synthetic data generator of Agrawal and Srikant does NOT produce realistic market basket data. Our analysis of the structures output by QUEST will be asymptotic in nature in that some of our results hold with probability tending to one as the size of the structure becomes large. The reader is referred to [5, 9] for the relevant graph-theoretic terminology, and to [11] for basic concepts in the Theory of Probability.

3.1 The model

We first review the model developed by Agrawal and Srikant. The synthetic data generator takes as parameters n, the number of different items, h, the number of transactions in the database, T, the average transaction size, and two other parameters I, and l whose meaning will become clear further on. It should however be said that at least two other quantities (referred to in the paper as *correlation level* ρ and *corruption level* c) and a number of other numerical values (kept hidden from the user) in the program affect its behaviour.

The program returns in fact two related structures: the actual database $\mathcal{D}$ and an auxiliary collection $\mathcal{T}$ of so called *(maximal) potentially large itemsets* or *patterns*, whose elements are used to populate $\mathcal{D}$. There are l itemsets in $\mathcal{T}$ and their average size is I.

The transactions in $\mathcal{D}$ are generated by first determining their size (picked from a Poisson distribution with mean equal to T) and then filling the transaction with the items contained in a number of itemsets selected independently and uniformly at random (u.a.r.) from $\mathcal{T}$.

Each itemset in $\mathcal{T}$ is generated by first picking the size of the itemset from a Poisson distribution with mean equal to I. Items in the first itemset are then chosen randomly. To model the phenomenon that large itemsets often have common items, some fraction (chosen from an exponentially distributed random variable with mean equal to the expected correlation level ρ) of items in subsequent itemsets are chosen from the previous itemset generated. The remaining items are picked at random. Each itemset in T has a weight associated with it, which corresponds to the probability that this itemset will be picked. This weight is picked from an exponential distribution with unit mean, and then normalised

so that the sum of the weights over all itemsets in T is 1. The next itemset to be put in the transaction is chosen from $\mathcal{T}$ by tossing an l-sided weighted coin, where the weight for a side is the probability of picking the associated itemset. To model the phenomenon that all the items in a large itemset are not always bought together, each itemset in $\mathcal{T}$ is assigned a corruption level c. When adding an itemset to a transaction, items from the itemset are dropped as long as a uniformly distributed random number between 0 and 1 is less than c.

Agrawal and Srikant claim that the resulting database $\mathcal{D}$ mimics a set of transactions in the retailing environment. Furthermore they justify the shape and the use of the structure $\mathcal{T}$ based on the fact that people tend to buy sets of items together, some people may buy only some of the items from a (potentially) large itemset, and on the other hand a transaction may contain more than one large itemset. Finally they observe that transaction sizes and the sizes of the large itemsets (although variable) are typically clustered around a mean and a few collections have many items.

The empirical analysis in Section 2 and the outcome of the mathematical investigation of the very crude approximation of QUEST described in the next Section indicate that perhaps some of these claims are a bit over-rated. Furthermore, in Section 4 we will argue that most of these assumptions are satisfied by a simpler model that also fits the real data more closely. In the next section, to support the claim that QUEST generated databases display an occurrence distribution that is quite different from that observed in real life market basket data, we analyse such distribution asymptotically in what could be considered to be a database generated by a simplified version of QUEST.

3.2 A simplified version of QUEST

In this Section we argue that, by its very mathematical properties, the QUEST generation process does not lead to datasets showing a heavy tailed item occurrence distribution.

To maximise the clarity of exposition we simplify the definition of the process by removing a number of features that wouldn't significantly change our main results, but would make the arguments more involved. Thus we will deal with pairs $(\mathcal{D}, \mathcal{T})$ where both $\mathcal{D}$ and $\mathcal{T}$ are defined on the same set of items $\mathcal{I}$. System parameters will be the following integers:

n the number of items in $\mathcal{I}$.

h the number of transactions (or *edges*) in $\mathcal{D}$.

l the number of patterns (or *edges*) in $\mathcal{T}$.

k the number of patterns per transaction.

s the size (the number of items) in each pattern

ρ the number of vertices shared between two consecutive patterns (with $\rho \in \{0, \ldots, s\}$).

Typically $h >> n$ (e.g. $h = n^{O(1)}$) while $l = O(n)$. Parameters k and s are small constants independent of n. Note that the assumption that the size of the transactions is variable as stated in Agrawal and Srikant's work is "simulated" by the fact that different patterns used to form a transaction may share some items. Following Agrawal and Srikant we use a correlation level ρ between subsequent patterns, but assume it takes the fixed constant value $s/2$. Hence $s/2$ items in each pattern (except the first one) belong its predecessor in the generation sequence and the remaining $s - \rho$ are chosen u.a.r. in $\mathcal{I}$. We do not use any corruption parameter.

We assume that $\mathcal{D}$ and $\mathcal{T}$ are populated by the following simplified version of Agrawal and Srikant's process.

1. Generate $\mathcal{T}$ by selecting the first pattern as a random set of size s over $\mathcal{I}$ and any subsequent pattern by choosing (with replacement) ρ elements u.a.r. from the last generated pattern and $s - \rho$ elements u.a.r. (with replacement) in $\mathcal{I}$.
2. Generate $\mathcal{D}$ by filling each transaction independently with the elements of k (not necessarily distinct) patterns in $\mathcal{T}$ chosen independently u.a.r.

Let $\mathrm{d}_{\mathcal{D}}(v)$ (resp. $\mathrm{d}_{\mathcal{T}}(v)$) be the random variable for the number of transactions in $\mathcal{D}$ (resp. patterns in $\mathcal{T}$) containing item v. In this context N_r will denote the random variable counting the number of items occurring in exactly r transactions of $\mathcal{D}$. Obviously items occurring in many patterns of $\mathcal{T}$ have a higher chance of occurring in many database transactions. The following result quantify the influence of $\mathcal{T}$ on the item occurrence distribution in $\mathcal{D}$.

THEOREM 1. *Let $(\mathcal{D}, \mathcal{T})$ with parameters n, h, l, k, s, and ρ as described above. Then*

1. *For each item $v \in \mathcal{I}$, the random variable $\mathrm{d}_{\mathcal{D}}(v)$ has binomial distribution with parameters h and*

$$p_{k,l} = \sum_{i=1}^{k} \binom{k}{i} (-1)^{i+1} \frac{\mathrm{E}(\mathrm{d}_{\mathcal{T}}(v))^i}{l^i},$$

   where $\mathrm{E}(\mathrm{d}_{\mathcal{T}}(v))^i$ is the i-th central moment of the random variable $\mathrm{d}_{\mathcal{T}}(v)$.

2. *$\mathrm{E}(N_r) = n \Pr[\mathrm{d}_{\mathcal{D}}(v) = r]$ and, furthermore, for any function $\omega(n)$ such that $\lim_{n\to\infty} \omega(n) = \infty$*

$$\Pr[|N_r - \mathrm{E}(N_r)| \geq \omega(n)\sqrt{n}] = O(\tfrac{1}{\omega(n)^2}).$$

PROOF. By definition the edges of $\mathcal{D}$ are generated independently of each other. An item v has degree r in $\mathcal{D}$ if it belongs to r fixed edges. Thus, in symbols:

$$\Pr[\mathrm{d}_{\mathcal{D}}(v) = r] = (\Pr[v \text{ belongs to a transaction}])^r$$

If we assume that each transaction is formed by the union of k patterns (chosen independently u.a.r. from $\mathcal{T}$) then $\mathrm{d}_{\mathcal{D}}(v)$ has binomial distribution. The probability that v belongs to a given transaction is:

$$\begin{aligned}\sum_{j=0}^{l}\Pr[\mathrm{Bin}(k,\tfrac{j}{l})\geq 1]\Pr[\mathrm{d}_{\mathcal{T}}(v)=j] &= \\ &= \sum_{j=0}^{l}(1-(1-\tfrac{j}{l})^k)\Pr[\mathrm{d}_{\mathcal{T}}(v)=j] \\ &= 1-\tfrac{1}{l^k}\sum_{j=0}^{l}(l-j)^k\Pr[\mathrm{d}_{\mathcal{T}}(v)=j] \\ &= 1-\tfrac{\mathrm{E}(l-\mathrm{d}_{\mathcal{T}}(v))^k}{l^k}\end{aligned}.$$

Now the result follows by the binomial theorem.

The second part of the Theorem statement follows easily from the first one, the independence assumptions and a direct application of Chebyshev's inequality. □

The distribution problem in $\mathcal{D}$ is thus reduced to finding the first k moments of the random variable $\mathrm{d}_{\mathcal{T}}(v)$. Solving the latter does not seem very easy in general. In the remainder of this Section we complete our analysis under more restricted assumptions.

3.2.1 Item occurrences in $\mathcal{T}$ when $s=2$

In this Section we look at the item occurrence distribution in $\mathcal{T}$ when $s=2$. It should be stressed that such restriction is not essential and is only kept for clarity of exposition. All techniques used to prove Theorem 2 and 3 below readily generalise to the analysis of $\mathcal{T}$ when $s>2$. Theorem 2 and 3 therefore are simplified versions of more general statements valid for collections of patterns $\mathcal{T}$ generated using a value of $s>2$ (modulo some obvious generalisation of some of the definitions given below).

If $s=2$ then $\mathcal{T}$ is a graph. The structure of such graph will depend on the value of ρ. If $\rho=0$ edges are chosen independently u.a.r. The resulting graph will be a variant of the well known fixed density model $\mathcal{G}(n,l)$ (see, for example, [13, Chp. 1]). At the other extreme if $\rho=2$ the resulting graph will only contain one edge (or, more precisely, l copies of the same edge). From now on we therefore focus on the case $\rho=1$.

When $s=2$ and $\rho=1$, the resulting graph can be thought of as directed, with edges chosen one after the other according to the following process:

1. The first directed edge e_1 is chosen as a random two element set in $\mathcal{I}$.
2. If the edge chosen as step i is $e_i=(w,z)$, (for $i\geq 1$), then e_{i+1} is chosen by:
   (a) selecting an item uniformly at random in $\mathcal{I}$ (each possible item may occur with probability $\frac{1}{n}$);
   (b) and then selecting the second element of the pair at random as either w or z (each choice occurs with probability $\frac{1}{2}$).

The degree of v in $\mathcal{T}$ can thus be seen as the sum of its in-degree $\mathrm{d}^-_{\mathcal{T}}(v)$ (number of edges having v as second component) and out-degree $\mathrm{d}^+_{\mathcal{T}}(v)$ (number of edges having v as first component). In what follow let $N_r^{\mathrm{sg}}=N_r^{\mathrm{sg}}(\mathcal{T})$ be the number of vertices having $\mathrm{d}^{\mathrm{sg}}_{\mathcal{T}}(v)=r$, for $\mathrm{sg}\in\{\text{“ ”},-,+\}$. Let $\delta^{\mathrm{sg}}(\mathcal{T})$ (resp. $\Delta^{\mathrm{sg}}(\mathcal{T})$) denote the smallest (largest) r such that $N_r^{\mathrm{sg}}\neq 0$. Also, let $p_r(\alpha)=\frac{\alpha^r\mathrm{e}^{-\alpha}}{r!}$ and $\sigma_r(\alpha)=p_r(\alpha)(1-p_r(\alpha)-p_r(\alpha)\frac{(\alpha-r)^2}{\alpha})$.

Theorem 2. *Let $\mathcal{T}$ be given with $s=2$, $\rho=1$ and all other parameters specified arbitrarily. Then*

1. *For each $v\in\mathcal{I}$, $\mathrm{d}^+_{\mathcal{I}}(v)$ has binomial distribution with parameters l and $\frac{1}{n}$.*
2. *If $\lim_{n\to\infty} l/n=\alpha<\infty$ then the probability distribution of N_r^+ tends to a normal distribution with parameters $np_r(\alpha)$ and $n\sigma_r(\alpha)$.*
3. *(Extrema) Under the assumptions above, $\Pr[\delta^+(\mathcal{T})=0]\to 1$. Furthermore it is possible to choose $r=O(\frac{\log n}{\log\log n})$ so that $r>\alpha$ and $\lim_{n\to\infty} np_r(\alpha)=\lambda<\infty$ and*

$$\Pr[\Delta^+(\mathcal{T})=r-1]\to \mathrm{e}^{-\lambda},\ \ \Pr[\Delta^+(\mathcal{T})=r]\to 1-\mathrm{e}^{-\lambda}.$$

Proof. Under assumptions of our model, the sequence $(\mathrm{d}^+_{\mathcal{T}}(v))_{v\in\mathcal{I}}$ has the same probability distribution as a vector describing the random allocation of l identical balls in n distinct urns. The result then follows from classical work on the subject (see for instance [14, Chp. I and II] □

The in-degrees can also be estimated through a slightly more elaborate argument. For each time step $i\geq 1$, let X_i be the second component of e_i. If $e_i=(w,z)$ and x is the first component of e_{i+1} then $X_{i+1}=w$ (resp. $X_{i+1}=z(=X_i)$) with probability $\frac{1}{2}$ and in that case $e_{i+1}=(x,w)$ (resp. $e_{i+1}=(x,z)$). Let $i_0=0$ and $1\leq i_1<\ldots<i_{\mathrm{d}^+_{\mathcal{T}}(v)}$ be the times when v is picked as starting-point of an edge (each occurrence of v as end-point comes into existence because, at some previous step, v was chosen as starting point of an edge). For $j\in\{0,\ldots,\mathrm{d}^+_{\mathcal{T}}(v)\}$ define $\zeta_j(v)$ as follows

$$\zeta_j(v)=\begin{cases}0 & X_{i_j+1}\neq v\\ \sum_{i_j+1\leq h<h_0}\mathbf{1}_{\{X_h=v\}} & \text{otherwise.}\end{cases}$$

In the expression above $h_0=\min\{h>i_j+1: X_h\neq v\vee h=i_{j+1}+1\}$. The function $\zeta_j(v)$ counts how many consecutive times v is the end-point of an edge from step i_j+1 to just before step h_0 (so, for example, if v is the first component of an edge at step $i_{j+1}=i_j+1$ then $\zeta_j(v)$ can be either 0 or 1). Then

$$\mathrm{d}^-_{\mathcal{T}}(v)=\sum_{j=0}^{\mathrm{d}^+_{\mathcal{T}}(v)}\zeta_j(v).$$

Note that for each j, $\zeta_j(v)\in\{0,\ldots,u_j\}$ with $u_j=i_{j+1}-i_j$ (where $i_{\mathrm{d}^+_{\mathcal{T}}(v)+1}=l$). Let $\mathcal{E}_v(d,i_1,\ldots,i_d)$ be the event "$\mathrm{d}^+_{\mathcal{T}}(v)=d$ and $i_1,\ldots,i_d$ being the time steps when v is selected u.a.r. as starting point of an edge". Conditioned

on $\mathcal{E}_v(d, i_1, \ldots, i_d)$, the variables $\zeta_j(v)$ are independent and have a truncated geometric distribution. We can use this information to get some estimates on the distribution of $\mathrm{d}^-_{\mathcal{T}}(v)$.

THEOREM 3. *Let $\mathcal{T}$ be given with $s = 2$, $\rho = 1$ and all other parameters specified arbitrarily. Then for each $v \in \mathcal{I}$ and $r \in \{0, \ldots, l\}$ there exists a function $L(r)$ such that*

$$\Pr[\mathrm{d}^-_{\mathcal{T}}(v) = r] = \frac{L(r)}{2^{r-l}} \sum_{d=0}^{l} (\frac{1}{n})^d (1 - \frac{1}{n})^{l-d} \sum_{i_1,\ldots,i_d} \prod_{j=0}^{d} \frac{1}{2^{u_j+1} - 1}.$$

PROOF. The probability that an item has in-degree r in $\mathcal{T}$ is

$$\sum_{d=0}^{l} (\frac{1}{n})^d (1 - \frac{1}{n})^{l-d} \sum_{i_1,\ldots,i_d} \Pr[\sum_{j=0}^{d} \zeta_j(v) = k \mid \mathcal{E}_v(d, i_1, \ldots, i_d)]$$

(where the term for $d = 0$ simplifies to $(1 - \frac{1}{n})^l \Pr[\zeta_0(v) = k \mid \mathrm{d}^+_{\mathcal{T}} = 0]$). Under the given conditioning, the $\zeta_j(v)$ are independent r.v.'s having right-truncated geometric distribution. The probability distribution of the sum of random variables of this type has been studied before [15]. It follows from this work that

$$\Pr[\sum_{j=0}^{d} \zeta_j(v) = r \mid \mathcal{E}_v(d, i_1, \ldots, i_d)] = \frac{L(r)}{2^{d+r+1}} \prod_{j=0}^{d} \frac{1}{1 - (\frac{1}{2})^{u_j+1}}$$

where $L(r)$ are defined by equations (11) and (12) in [15]. The result follows. □

3.2.2 The occurrence distribution in $\mathcal{D}$ in a very simple case

In this Section we further simplify our model, assuming that $k = 1$, i.e. each transaction in $\mathcal{D}$ is formed by a single random edge of $\mathcal{T}$.

The following result shows that, when n becomes large, in such simple setting the item occurrence distribution decays super-polynomially (and therefore it cannot have, asymptotically, a heavy tail).

THEOREM 4. *For $k = 1$, $N_r = o(r^{-\alpha})$ for any fixed positive α.*

PROOF. If $k = 1$ by Theorem 1 $\mathrm{d}_{\mathcal{D}}(v)$ has binomial distribution. We claim that

$$\tfrac{3}{2}(\mathrm{E}(\mathrm{d}^+_{\mathcal{T}}(v)) + \tfrac{1}{3}) \leq \mathrm{E}(\mathrm{d}_{\mathcal{T}}(v)) \leq 2\mathrm{E}(\mathrm{d}^+_{\mathcal{T}}(v)) + 1.$$

From this we have that, for $l \to \infty$,

$$\tfrac{3}{2n} + o(1) \leq p_{1,l} \leq \tfrac{2}{n} + o(1),$$

and therefore, asymptotically, the binomial distribution of $\mathrm{d}_{\mathcal{D}}(v)$ is sandwiched between two Poisson distributions. the result follows using linearity of expectation and the last part of Theorem 2.

To believe the claim notice that, for each v, $\mathrm{d}_{\mathcal{T}}(v) = \mathrm{d}^+_{\mathcal{T}}(v) + \mathrm{d}^-_{\mathcal{T}}(v)$, hence, by linearity of expectation and Theorem 2

$$\mathrm{E}(\mathrm{d}_{\mathcal{T}}(v)) = \tfrac{l}{n} + \mathrm{E}(\mathrm{d}^-_{\mathcal{T}}(v)).$$

Conditioning on $\mathcal{E}_v(d, i_1, \ldots, i_d)$ we can write, for any $k \geq 1$,

$$\mathrm{E}(\mathrm{d}^-_{\mathcal{T}}(v)^k) = \mathrm{E}(\mathrm{E}(\mathrm{d}^-_{\mathcal{T}}(v)^k \mid \mathrm{d}^+_{\mathcal{T}}(v), I_1, \ldots, I_{\mathrm{d}^+_{\mathcal{T}}(v)}))$$

(where I_j is the random variable equal to i if the jth occurrence of v as first element of an edge is at step i). By the assumptions on the $\zeta_j(v)$, for any $j \in \{0, \ldots, \mathrm{d}^+_{\mathcal{T}}(v)\}$,

$$\Pr[\zeta_j(v) = k \mid \mathcal{E}_v(d, i_1, \ldots, i_d)] = \begin{cases} (\frac{1}{2})^{k+1} & 0 \leq k < u_j \\ (\frac{1}{2})^{u_j} & \text{otherwise} \end{cases}.$$

Furthermore, by the independence assumption, we can easily compute the moment generating function of $\mathrm{d}^-_{\mathcal{T}}(v)$ (conditional on $\mathcal{E}_v(d, i_1, \ldots, i_d)$):

$$M_{\mathcal{E}}(t) = (\mathrm{e}^t - 2)^{-(d+1)} \textstyle\prod_{j=0}^{d} ((\frac{\mathrm{e}^t}{2})^{u_j} (\mathrm{e}^t - 1) - 1).$$

From this, through differentiation,

$$\mathrm{E}(\mathrm{d}^-_{\mathcal{T}} \mid \mathcal{E}_v(d, i_1, \ldots, i_d)) = d + 1 - \textstyle\sum_{j=0}^{d} (\frac{1}{2})^{u_j},$$

and the claim follows since, obviously,

$$\tfrac{d+1}{2^{l/(d+1)}} \leq \textstyle\sum_{j=0}^{d} (\frac{1}{2})^{u_j} \leq \tfrac{d+1}{2}$$

□

We close this section by noticing another peculiar feature of QUEST. For $k = 1$, there is a constant probability that a given item will never occur in a transaction of $\mathcal{D}$. Equivalently a constant fraction of the n available items will never occur in the resulting database. This phenomenon was observed in the two synthetic databases analysed in Section 2: `T40I10D100K.dat` only uses 941 of the 1,000 available items, `T10I4D100K.dat`, only 861. Of course this irrelevant from the practical point of view, but it's a strange artifact of the choice of having a two-component structure in the QUEST generator.

4. AN ALTERNATIVE PROPOSAL

In this section we put forward an alternative model for generating synthetic databases. Our model is in line with the proposal of Barabási and Albert [3] (originally introduced to model structures like the scientific author citation network or the world-wide web). The proofs of our main result mimics that of similar results presented by Cooper [8].

The model contains a mechanism (called *preferential attachment*) that allows the process that generates the transactions in $\mathcal{D}$ one after the other to choose their components based on the frequency distribution of such items in previously generated transactions. Such mechanism seems to be necessary to generate datasets that are closer to the real ones analysed in Section 2.

Instead of assuming an underlying set of patterns $\mathcal{T}$ from which the transactions $\mathcal{D}$ are built up, the transactions are generated sequentially. Items which have already occurred in previous transactions are called *old*, whereas items occurring for the first time in the current transaction are called *new*. At the start there is an initial set of e_0 transactions on n_0 existing (old) items. The model can generate transactions based entirely on the n_0 old items, but in general we assume that new items can also be added during transactions, so that at the end of the simulation the total number of items is $n > n_0$. The simulation proceeds for t steps generating groups of transactions (which can be of size 1 if desired) at each step. For each transaction (or group of transactions) in the sequence there are four choices made by the simulation:

(i) The type of transaction OLD or NEW. An OLD type transaction consists entirely of items occurring in previous transactions. A NEW type transaction consists of a mix of new items and items occurring in previous transactions. For simplicity we assume each NEW transaction adds one new item, and at least one old item.

(ii) The number of transactions in the group. This can be fixed at one transaction, or given any required probability distribution. Grouping corresponds to e.g. the persistence of a particular item in a group of transactions in the QUEST model.

(iii) The size of the individual transactions. This can be fixed, or given a probability distribution.

(iv) The method of choosing the items in the transaction. This is assumed to be a mix of u.a.r (items chosen randomly) and preferential attachment (reflects the popularity of the items in previous transactions).

This outlines the general model. The way that these choices affect the degree distribution and its form are described below. Our main result is that provided the number of transactions is large, the distribution of occurrences of items in the transaction sample follows a power law, for which the parameter z, is given by $z = 1 + 1/\eta$, where η is the (expected) proportion of items selected by preferential attachment in all transactions. That is to say, $N_r = N_r(t)$ the number of items which occur r times in t transactions satisfies $N_r \sim Ctr^{-z}$ for large r and some constant C.

We assume that some preferential attachment choices will occur, and our model allows for the case where all choices are made preferentially. In order to give the exact value of η arising from (i)-(iv) above we now describe the model in detail.

For convenience of description we call items *vertices* and transactions *edges* or *hyper-edges*. The set of transactions at any step is viewed as a (hyper-)graph.

The edge (transaction) sizes are fixed or given by a distribution. Let $\boldsymbol{\pi} = (\pi_2, ..., \pi_r)$ be the distribution of edge sizes $j = 2, ..., r$. For such $\boldsymbol{\pi}$, the average edge (transaction) size is $\overline{\pi} = \sum j\pi_j$.

There is an initial set of transactions defining a graph $G(0)$ of finite size: n_0 items and e_0 edges. Loops and multiple edges are allowed in $G(0)$ and can be formed in $G(t)$ at any step t. The cardinality of the vertex set, edge set and the total degree (total number of items purchased by all customers) is $V(t), E(t), D(t)$ respectively.

At each step $t = 1, 2, ...$, of the simulation an independent choice is made between a NEW and an OLD procedure. Let α be the probability the NEW procedure is followed and $\beta = 1 - \alpha$ the probability the OLD procedure is followed. If we assume $\alpha > 0$, the expected size of the vertex set increases with t. That is to say, new items are added to the set of purchase transactions as time goes on.

For NEW edges we assume that there is one new vertex (new item) in the edge, and the other vertices (old items) are chosen independently with probability p_A. For OLD edges, we assume that all vertices (items) are chosen independently with probability p_C. The values of p_A, p_C (which can be the same) are given below.

NEW **procedure.** Chosen with probability α at step t. A new vertex v_t is added. There are $m(t)$ edges directed to $G(t-1)$. Thus $m(t)$ is the size of the transaction group. Each edge in the transaction group has its own size sampled from $\boldsymbol{\pi}$. If we wish we can set $m(t) = 1$ or some other constant value or allow it to assume a distribution.

In general, let p_m denote the probability $\Pr[m(t) = m]$ that m edges are added to v_t at step t. For convenience we refer to vertex v_t by its step label t (although not every step will necessarily have a corresponding vertex, ie new item). Thus the number $m(t)$ of edges originating from v_t is a random variable. We assume $m(t) \geq 1$, and that $m(t)$ is sampled from a distribution $\boldsymbol{p} = (p_1, ..., p_m, ..., p_L)$. Let $\overline{m} = \mathrm{E}(m(t))$. We have that $\overline{m} \geq 1$. The degree $m(t)$ of vertex v_t after it is added is denoted by $d(t, t)$. Similarly $d(v, t)$ is the degree of vertex v after step t.

Each new edge $e_i, i = 1, ..., m$ makes an independent decision to select its terminal vertex w in $G(t-1)$ either by preferential attachment or u.a.r. with probabilities $A_1,\ A_2 = 1 - A_1$ respectively. Thus the probability that w is selected by e_i is

$$p_A(w, t) = A_1 \frac{d(w, t-1)}{D(t-1)} + A_2 \frac{1}{V(t-1)}, \qquad (1)$$

where $d(w, t-1)$ is the degree of w at the end of step $t-1$.

OLD **procedure.** Chosen with probability $\beta = 1 - \alpha$ at step t. Insert $M(t)$ edges into $G(t-1)$ according to a distribution $\boldsymbol{q} = (q_1, ..., q_M, ..., q_L)$. Let $\overline{M} = \mathrm{E}(M(t))$. The size of each edge is given by the distribution $\boldsymbol{\pi}$. The choice of vertices (items) of each inserted edge is made independently with

probability p_C from u.a.r./preferential rules similar to (1), but with parameters C_1, C_2. Thus we have

$$p_C(v,t) = C_1 \frac{d(v,t-1)}{D(t-1)} + C_2 \frac{1}{V(t-1)}. \quad (2)$$

We now give the theorem for the (asymptotic) power law distribution on the items in the transaction sample, and discuss some special cases.

THEOREM 5. *As the number of items $n \to \infty$, with probability tending to 1, the value of η is*

$$\eta = \frac{\alpha\overline{m}(\overline{\pi}-1)A_1 + \beta\overline{M}\overline{\pi}C_1}{(\alpha\overline{m}+\beta\overline{M})\overline{\pi}}.$$

For fixed value of t, the expected values of $V(t), E(t), D(t)$ are

$$\mathrm{E}(V(t)) = n_0 + \alpha t,$$

$$\mathrm{E}(E(t)) = m_0 + t(\alpha\overline{m} + \beta\overline{M}),$$

$$\mathrm{E}(D(t)) = d_0 + t((\alpha\overline{m} + \beta\overline{M})\overline{\pi}).$$

PROOF. The proof proceeds by deriving the degree distribution $d(v,t)$ of each vertex v, excepting the small fraction of vertices occurring early in the process.

The distribution of $X(t)$, the number of edges selecting vertex v at step t, is given by

$$X(t) \sim \mathbf{1}_\alpha(t) \left(\textstyle\sum_{i=1}^{m(t)} \mathrm{Bin}(R_i^\pi - 1, p_A(t))\right)$$

$$+\mathbf{1}_\beta(t) \left(\textstyle\sum_{i=1}^{M(t)} \mathrm{Bin}(R_i^\pi, p_C(t))\right)$$

where $R_i^\pi \sim \boldsymbol{\pi}$ is a random variable giving the size of edge $e_i,\ i = 1, ..., m(t)$ in the NEW procedure, and of edge $e_i,\ i = 1, ..., M(t)$ in the OLD procedure, and $\mathrm{Bin}(\ldots)$ denote arbitrary random variables having binomial distribution with the specified parameters.

Let η be defined as above, let ν be defined by

$$\nu = \frac{\alpha\overline{m}(\overline{\pi}-1)A_2 + \beta\overline{M}\overline{\pi}C_2}{\alpha},$$

and let

$$\xi(v) = m(v) + \nu/\eta.$$

By considering $X(t)$ within the set $I(t)$ of good histories, $(V(t) = (1+o(1))\mathrm{E}(V(t)),\ E(t) = (1+o(1))\mathrm{E}(E(t)))$ we find that

$$\Pr[X(t+1) = 0 \mid d(v,t) = m+j,\ I(t)] =$$
$$= 1 - \frac{\eta(m+j)+\nu}{t}\left(1+O(\mathrm{e}(t,j))\right)$$
$$\Pr[X(t+1) = 1 \mid d(v,t) = m+j,\ I(t)] =$$
$$= \frac{\eta(m+j)+\nu}{t}\left(1+O(\mathrm{e}(t,j))\right),$$

etc where

$$\mathrm{e}(\tau, j) = \sqrt{\log\tau/\tau} + (m+j)/\tau.$$

Let v, t be fixed, suppose $d(v,t) = m+l$ and let $\mathbf{T} = (T_j,\ j = 1, ..., l)$ give the steps T_j (if any) at which the degree of v changed. Let $\boldsymbol{\tau} = (\tau_1, ..., \tau_l)$ denote a particular value of $\mathbf{T}$, so that τ_j is the step at which $d(v,\tau_j)$ changed from $m+j-1$ to $m+j$ in this case. If more than one edge hit v during step τ_j, the value τ_j is repeated the appropriate number of times in $\boldsymbol{\tau}$. We define $\tau_0 = v$ and $\tau_{l+1} = t$. For $v < \tau \le t$ let $J = \{\boldsymbol{\tau} : \tau_1 \le \tau_2 \le \cdots \le \tau_l\}$ be the sequences of possible transitions. Thus

$$\Pr[d(v,t) = m+l \mid d(v,v) = m] = \sum_{\boldsymbol{\tau}\in J} \Pr[\mathbf{T} = \boldsymbol{\tau}].$$

The proof proceeds by evaluating the sum over the possible sequences $\boldsymbol{\tau} \in J$, using $\Pr[X(t) = k]$, $k = 0, 1, 2, ...$ given above to establish that

$$\Pr[d(v,t) = m+l] =$$
$$= (1+o(1)) \binom{l+\xi-1}{l} \left(\frac{v}{t}\right)^{\eta\xi} \left[1 - \left(\frac{v}{t}\right)^{\eta}(1+o(1))\right]^l$$
$$+ O(v^{-K}),$$

where $o(1) = 1/(\log t)^2$.

By fixing m, l and summing over v we can now estimate the proportion of vertices of a given degree. Let

$$n_m(l) = \frac{\Gamma(\xi+1/\eta)}{\eta\Gamma(\xi)} \frac{\Gamma(l+\xi)}{\Gamma(l+\xi+1+1/\eta)}, \quad (3)$$

where $\Gamma(a) = (a-1)\Gamma(a-1)$. For $l \ge 0$ define $N_{l,m}(t)$ as the number of vertices of $\mathcal{I}$ with $d(v,v) = m$ and $d(v,t) = l+m$. Recall that p_m is the probability that $d(v,v) = m$ and let $\omega = \log t$.

For $\eta < 1/2$ let $l^\star = t^{\eta/3}$, and for $\eta \ge 1/2$ let $l^\star = t^{1/6}/\omega^2$. For $0 \le l \le l^\star$ we have:

(a) Expected degree sequence

$$\mathrm{E}(N_{l,m}(t)) = \alpha p_m n_m(l) t \left(1 + O\left(\tfrac{1}{\log t}\right)\right).$$

(b) Concentration

$$\Pr\left[|N_{l,m}(t) - \mathrm{E}(N_{l,m}(t))| \ge \frac{\mathrm{E}(N_{l,m}(t))}{\sqrt{\log t}}\right] = O\left(\frac{1}{\log t}\right).$$

To get the asymptotics, we note from (3) that

$$n_m(l) = \left(1 + O\left(\tfrac{1}{l}\right)\right) \tfrac{\Gamma(\xi+1/\eta)}{\eta\Gamma(\xi)}\, l^{-\left[1+\frac{1}{\eta}\right]},$$

and as $l \to \infty$, $N_{l,m}(t) \sim Ctl^{-(1+1/\eta)}$, a power law with parameter $1 + 1/\eta$. Details of the process are hidden in the model specific constant, $C = C(\eta, \nu, m) = \alpha p_m \Gamma(m + \frac{\nu+1}{\eta})/(\eta\Gamma(m+\frac{\nu}{\eta}))$. □

We next summarize the parameters of our model, and compare them with those of the QUEST model. Our model generates groups of transactions for t steps resulting in a set $\mathcal{D}$ of transactions. In expectation there are $n = n_0 + \alpha t$ items in the set $\mathcal{I}$ and $h = e_0 + t(\alpha\overline{m} + \beta\overline{M})$ transactions, where $\overline{m}$, (resp. $\overline{M}$) are the average group sizes of new (resp. old) transactions generated at any step. The expected size of a

transaction is $\overline{\pi}$. The number of edges $m(t)$ in a group with a new vertex v_t, has the same function as the parameter ρ of QUEST, giving the persistence of a new item in the group of transactions.

There is no direct correspondence in our model to the parameters l, k, s of the QUEST model, as there are no *patterns* as such. Rather, consumers express their purchasing preferences as a function of the history of the (simulated) process in an adaptive way, rather than having all the history (patterns) determined a priori. If desired, it would be possible to introduce pattern behaviour in our model as follows: The existing items (or a subset of them) would be designated a type $i = 1, ..., l$. When an edge is generated, it would sample a subset of its vertices from the designated types. As an extreme example, we could set the edge size r to be $r = ls$ where l is the number of types, and s is the pattern size of the type, and then s vertices of each type would be sampled for the edge.

Turning to examples, in the simplest case, the group sizes are fixed $m(t), M(t) = m$ for example $m = 1$, and the preferential attachment behaviour of the transaction types is the same $A_1 = C_1$. Thus

$$\eta = 1 - \frac{\alpha A_1}{\overline{\pi}}, \qquad z = 1 + \frac{\overline{\pi}}{\overline{\pi} - \alpha A_1},$$

where z is the power law parameter, α is the probability of a new item in a transaction, A_1 is the probability the item is chosen preferentially and $\overline{\pi}$ is the average transaction size. This can be further simplified by setting $A_1 = 1$ (all behaviour is preferential attachment: we are strict followers of fashion). If we allow $\alpha \to 0$, the case where the initial graph $G(0)$ determines everything, and no new items are added, then $z \to 2$.

5. REFERENCES

[1] R. C. Agarwal, C. C. Aggarwal, and V. V. V. Prasad. Depth first generation of long patterns. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 108–118, New York, NY, USA, 2000. ACM Press.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. Expanded version available as IBM Research Report RJ9839, June 1994.

[3] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[4] R. J. Bayardo. Efficiently mining long patterns from databases. In *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 85–93, New York, NY, USA, 1998. ACM Press.

[5] C. Berge. *Graphs and Hypergraphs*, volume 6 of *North-Holland Mathematical Library*. North-Holland, 1973.

[6] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets. Using association rules for product assortment decisions: a case study. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 254–260, New York, NY, USA, 1999. ACM Press.

[7] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD '97: Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 255–264, New York, NY, USA, 1997. ACM Press.

[8] C. Cooper. The age specific degree distribution of web-graphs. *Combinatorics, Probability and Computing*, 15(5):637–661, 2006.

[9] R. Diestel. *Graph Theory*, volume 173 of *Graduate Texts in Mathematics*. Springer-Verlag, 1999.

[10] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *ACM SIGCOMM Computer Communication Review*, 29(4):251–262, 1999.

[11] W. Feller. *An Introduction to Probability Theory and its Applications*. John Wiley and Sons, 1968.

[12] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 1–12, New York, NY, USA, 2000. ACM Press.

[13] S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs*. John Wiley and Sons, 2000.

[14] V. F. Kolchin, B. A. Sevast'yanov, and V. P. Chistyakov. *Random Allocations*. V. H. Winston and Sons, 1978.

[15] G. Lingappaiah. Distribution of the sum of independent right truncated geometric variables. *Statistica*, 51(3):411–421, 1991.

[16] B. Mandelbrot. The Pareto-Levy law and the distribution of income. *International Economic Review*, 1:79–106, 1960.

[17] P. W. Purdom, D. Van Gucht, and D. P. Groth. Average-case performance of the apriori algorithm. *SIAM Journal on Computing*, 33(5):1223–1260, 2004.

[18] S. Redner. How popular is your paper? an empirical study of the citation distribution. *European Physical Journal*, B 4:401–404, 1998.

[19] A. Savasere, E. Omiecinski, and S. B. Navathe. An efficient algorithm for mining association rules in large databases. In *VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases*, pages 432–444, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[20] M. Steyvers and J. B. Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29:41–78, 2005.

[21] H. Toivonen. Sampling large databases for association rules. In *VLDB '96: Proceedings of the 22th International Conference on Very Large Data Bases*, pages 134–145, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc.

[22] D. J. Watts. The "new" science of networks. *Annual Review of Sociology*, 30:243–270, 2004.

[23] U. Yule. A mathematical theory of evolution based on the conclusions of Dr. J. C. Wills, F. R. S. *Philosophical Transactions of the Royal Society of London*, 213 B:21–87, 1925.

[24] M. J. Zaki and M. Ogihara. Theoretical foundations of association rules. In *Proc. 3rd SIGMOD Worksh. Research Issues in Data Mining and Knowledge Discovery*, pages 7:1–7:8, 1998.

[25] Z. Zheng, R. Kohavi, and L. Mason. Real world performance of association rule algorithms. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 401–406, New York, NY, USA, 2001. ACM Press.

[26] G. K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, 1949.

[27] O. Zaïane, M. El-Hajj, Y. Li, and S. Luk. Scrutinizing frequent pattern discovery performance. In *ICDE '05: Proceedings of the 21st International Conference on Data Engineering (ICDE'05)*, pages 1109–1110, Washington, DC, USA, 2005. IEEE Computer Society.