

# Empirical Bayes Screening for Multi-Item Associations

William DuMouchel and Daryl Pregibon

AT&T Labs—Research

180 Park Avenue

Florham Park, NJ 07932, USA

{dumouchel, daryl}@research.att.com

## ABSTRACT

This paper considers the framework of the so-called “market basket problem”, in which a database of transactions is mined for the occurrence of unusually frequent item sets. In our case, “unusually frequent” involves estimates of the frequency of each item set divided by a baseline frequency computed as if items occurred independently. The focus is on obtaining reliable estimates of this measure of interestingness for all item sets, even item sets with relatively low frequencies. For example, in a medical database of patient histories, unusual item sets including the item “patient death” (or other serious adverse event) might hopefully be flagged with as few as 5 or 10 occurrences of the item set, it being unacceptable to require that item sets occur in as many as 0.1% of millions of patient reports before the data mining algorithm detects a signal. Similar considerations apply in fraud detection applications. Thus we abandon the requirement that interesting item sets must contain a relatively large fixed minimal support, and adopt a criterion based on the results of fitting an empirical Bayes model to the item set counts. The model allows us to define a 95% Bayesian lower confidence limit for the “interestingness” measure of every item set, whereupon the item sets can be ranked according to their empirical Bayes confidence limits. For item sets of size  $J > 2$ , we also distinguish between multi-item associations that can be explained by the observed  $J(J-1)/2$  pairwise associations, and item sets that are significantly more frequent than their pairwise associations would suggest. Such item sets can uncover complex or synergistic mechanisms generating multi-item associations. This methodology has been applied within the U.S. Food and Drug Administration (FDA) to databases of adverse drug reaction reports and within AT&T to customer international calling histories. We also present graphical techniques for exploring and understanding the modeling results.

## Categories and Subject Descriptors

H.2.8 [Database Management] Database Applications - *data mining, statistical databases.*

## General Terms

Statistical Models, Data Mining, Knowledge Discovery.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.  
KDD 01 San Francisco CA USA  
Copyright ACM 2001 1-58113-391-x /01/08...\$5.00

## Keywords

Association rules, empirical Bayes methods, gamma-Poisson model, market basket problem, shrinkage estimation.

## 1. THE MARKET BASKET PROBLEM

A common data mining task is the search for associations of items in a database of transactions. Each transaction, such as a list of items purchased, or a list of diagnoses and medications in a patient medical report, defines a subset of all possible items in the union of all transactions. If there are  $K$  total items, we define for each transaction, binary random variables  $X_1, \dots, X_K$  that are 1 if the corresponding item is included in the transaction, and 0 otherwise. If the total number of transactions in the database is  $N$ , then the data being modeled could be represented as an  $N \times K$  matrix  $X = X_{ik}$ ,  $i = 1, \dots, N$ ;  $k = 1, \dots, K$ . To add more realism to the statistical model, we do not assume that the distribution of each row of  $X$  is identical. Rather, we assume that the transactions are stratified, for example by store location in the case of supermarket scanner data, or by patient age, sex, etc. in the case of medical data. Time period of the transaction could also be a stratification variable. The purpose of stratification is to avoid finding spurious associations between two items merely because they both tend to be more frequent in the same strata. We assume that associations due to stratification variables (so called *between* strata associations) are not of interest, and instead require the methodology to identify items that are associated *within* strata. For example, certain upscale store locations might have greater sales volume of high-priced items, but we may only wish to say such items have interesting associations if they tend to be purchased together more frequently than their per-store marginal frequencies indicate, since an overall association would show up even if the items were purchased independently within each store.

As the preceding paragraph indicates, we are departing from the majority of published treatments of the market basket problem by going beyond the enumeration of the support (proportion of transactions) of frequent item sets. No matter how frequently an item set occurs, we assume that it is of little interest to the analyst if its frequency is about the same as would be predicted if the members of the item set occur independently within each stratum. This comparison frequency, based on within-strata independence, will be called the *baseline* frequency of the particular item set, and our measure of interestingness is the ratio of actual to baseline frequency. Two earlier papers that focus on interestingness measures as deviations from expected frequencies predicted by independence for the market basket problem are Silverstein, Brin and Motwandi [14] and Aggarwal and Yu [1]. Considering the former paper first, [14] rightly criticizes the more common support-confidence framework for association rules and suggests using a chi-squared test statistic

for the hypothesis of complete independence of items in an item set as the basis for comparing the degree of dependence among different item sets. They develop a theory for using this test statistic that relies on the property of *upward closure*, namely that the computed chi-squared statistic testing independence for a set of items (binary variables in a database) can only increase if you increase the number of items (size or dimensionality of the item set). Unfortunately, the authors mistakenly state that the degrees of freedom of the chi-squared statistic is always just 1, for all values of  $J$ , the item set size. In fact, the degrees of freedom of their test statistic is  $2^J - J - 1$ . Thus, although the test statistic will increase as more variables are included in the computations, the statistical significance or “p-value” will not be monotone and may increase or decrease, thereby invalidating the supposed theoretical properties of their proposed procedures and computational methods. In addition, as demonstrated in DuMouchel [8] and discussed in Section 3 below, relying solely on a statistical significance test statistic for measuring dependence has the built-in flaw of being overly sensitive to sample size. The latter paper [1] also criticizes the standard support-confidence framework for item sets, pointing out how it is susceptible to identifying spurious associations. These authors propose a measure they call *collective strength*. If  $I$  is an item set, let  $v(I)$  be the *violation rate* of the item set in the database, which is the fraction of transactions that contain some, but not all, of the members of  $I$ . Then, if  $E[\cdot]$  refers to expectation under the assumption of independence of items, collective strength is defined as

$$C(I) = (1 - v(I)) E[v(I)] / v(I) (1 - E[v(I)])$$

For example, for item sets of size 3, the notion of collective strength pools the proportions of the two triples (0,0,0) and (1,1,1), and contrasts their proportion with  $v(I)$ , the pooled proportions of the other six possible triples. Then the ratio  $(1 - v(I))/v(I)$  is divided by the corresponding ratio of expected values under independence. Aggarwal and Yu [1] show that collective strength avoids the undesirable spuriousness to which the support-confidence framework is susceptible, yet has at least as nice computational properties, which allow the efficient enumeration of what they define as “strongly collective item sets.” However, in our opinion, this is achieved at a considerable loss of interpretability and practical value as a measure of association. For example, let us continue the example of an item set of size 3, and suppose the three items each occur in 10% of transactions. Under the independence assumption, the probabilities of observing (0,0,0) and (1,1,1) are  $.9^3 = .729$  and  $.1^3 = .001$ , respectively, leading to  $E[1 - v(I)] = .729 + .001 = .73$  and  $E[v(I)] = .27$ . Note how the probability of observing (0,0,0) dominates this calculation, compared to that of (1,1,1), even though it is the simultaneous presence (not absence) of the three items that is almost always of more interest in applications. Suppose that the observed probabilities of (0,0,0) and (1,1,1) are .75 and .02, in which case the triple of items is present 20 times more than expected from independence. But this sense of a 20:1 increase is lost in the computation of collective strength, where  $C(I) = (.77)(.27) / (.23)(.73) = 1.24$  is close to 1, the value of  $C(I)$  under independence. This undesirable property of collective strength will persist whenever the individual items are not very frequent (e.g., when most items occur in less than one-third of

transactions) and we believe this describes the majority of market basket problems.

Both the cited papers, as well as [6], refer to the simple ratio of observed to baseline frequencies of occurrence as the *interest* of an item set, which we denote  $R$ . The primary criticisms of  $R$  as a measure of dependence are

1. Since  $R$  is bounded above by  $1/P(A)P(B)\dots$ , if all the items  $A, B, \dots$  in an item set are quite probable, the range of  $R$  can be restricted to seemingly small values.
2. Since  $R$  is symmetric in the items  $A, B, \dots$ , it does not indicate predictive power or suggest causation.
3. Computing all item sets with greater than a specified value of  $R$  will be costly computationally, since a great many item sets with small support can have large values of  $R$ .

We believe that the first criticism is mitigated by the fact that the vast majority of market basket problems deal with items whose probabilities are not large. Data mining is, by definition, a search for low probability events of high value. We view the lack of directionality of  $R$  as more an advantage than a disadvantage. Once an item set is identified as unusually frequent, domain knowledge is required to identify mechanisms and potential causality among the items in the item set, something far beyond the presumption of a statistical algorithm. The third criticism above identifies both a computational and a statistical issue. We do not focus on computational issues here, which are already well treated in the association rule literature, but rather we focus on the statistical issues that arise for item sets with small support.

A novel element of our approach to identifying interesting item sets is the ability to detect interesting associations even if they have relatively small support. To do so, we must explicitly allow for noise or spurious association due to small samples. Consider an item set observed in 200 transactions, even though the baseline (independence model) frequency is just 2. How should this 200:2 ratio of actual to baseline be compared to other item sets having the same ratio, such as 20:0.2, or even 2:0.02? Our answer involves an empirical Bayes model, described in detail in [8], that produces a posterior distribution for the true ratio of actual to baseline expected frequencies, assuming that the observed counts of every item set have separate Poisson distributions. Our conservative estimate of the ratio of actual to baseline is the lower 95% Bayesian confidence limit for this ratio, which we denote  $\lambda_{.05}$ . When observed counts are large,  $\lambda_{.05}$  will be almost identical to the naïve observed ratio, namely  $R = 100$  in the 200:2 example. However,  $\lambda_{.05}$  will be rather less than 100 for the case of 20:0.2, and will be much smaller, perhaps near 1, in the 2:0.02 example. This tendency of empirical Bayes estimates to reduce the observed effect when sample sizes are small gives rise to the label *shrinkage estimate*.

To summarize, we introduce three improvements to the existing market basket methodology. First, we allow stratification adjustments to eliminate spurious association due to non-interesting stratification variables. Second, we focus on the ratio of item set frequencies to baseline frequencies computed via an independence assumption, rather than merely the support of each item set. And third, we replace the observed ratios of actual to baseline frequencies by empirical Bayes shrinkage estimates in

order to smooth away the noise due to small samples. A fourth innovation, applicable only to item sets of size greater than 2, will be described in Section 4.

## 2. BASELINE AND RELATIVE RATES

Let  $N_s$  be the number of transactions in stratum  $s$ , let  $N_{sj}$  be the number of transactions in stratum  $s$  that include item  $j$ , let  $N_{sjk}$  be the number of transactions in stratum  $s$  that include both item  $j$  and item  $k$ , and so forth, where  $s = 1, \dots, S$ , the number of strata, and  $1 \leq j < k \leq K$ . Define  $P_{sj} = N_{sj}/N_s$  as the proportion of transactions in stratum  $s$  that include item  $j$ . The baseline frequencies for item sets of size 1, 2, 3, and 4 are, respectively,  $e_j$ ,  $e_{jk}$ ,  $e_{jkl}$ , and  $e_{jklm}$ , where

$$\begin{aligned} e_j &= \sum_s N_s P_{sj} = \sum_s N_{sj} & e_{jk} &= \sum_s N_s P_{sj} P_{sk} \\ e_{jkl} &= \sum_s N_s P_{sj} P_{sk} P_{sl} & e_{jklm} &= \sum_s N_s P_{sj} P_{sk} P_{sl} P_{sm} \end{aligned}$$

Analogous definitions are used for baseline frequencies of larger item sets. This method of computing separate baseline frequency estimates for each stratum and then adding across strata became standard in the biostatistics literature after the article by Mantel and Haenzel [12]. The method provides protection against *Simpson's paradox* [15] whereby associations that hold separately within each stratum can disappear or even be reversed when data are pooled across strata. DuMouchel [8, Table 5 and related discussion] provides an example of how failure to stratify by age could lead to an incorrect assessment that polio vaccine is associated with sudden infant death syndrome (SIDS). Since polio vaccine is overwhelmingly given to infants, and of course only infants develop SIDS, a database including suspected adverse drug reactions for patients of all ages will show an excess of SIDS cases from polio vaccine recipients, although many well-regarded clinical trials of vaccine recipients show no such relationship.

We use lower case  $ns$  for total across-stratum counts:  $n_j = \sum_s N_{sj} = e_j$ ,  $n_{jk} = \sum_s N_{sjk}$ , etc. Then, for all item sets defined by  $(j < k)$ ,  $(j < k < l)$ ,  $(j < k < l < m)$ , etc., the raw *relative rates* are defined as

$$R_{jk} = n_{jk}/e_{jk}, \quad R_{jkl} = n_{jkl}/e_{jkl}, \quad R_{jklm} = n_{jklm}/e_{jklm}.$$

In domains for which only very large counts (e.g.,  $n > 500$ ) are of interest, then sorting and otherwise examining item sets according to their values of  $R$  may provide a sufficient analysis. It is recommended that separate examinations be conducted for item sets of size 2, of size 3, etc., because in general the independence baseline frequencies tend to fit the raw  $ns$  more poorly for larger item sets than for smaller item sets. That is, the distribution of values of  $R$  will usually be different for pairs than for triples or quadruples or higher-way item sets.

## 3. EMPIRICAL BAYES ESTIMATION

The empirical Bayes approach smoothes the values of  $R$  to allow more reliable estimates for smaller sample sizes. The empirical Bayes process is repeated separately for each size item set. The process for pairs is the same as for triples, quadruples, etc., so for simplicity we will drop the subscripts in our notation and assume that we have a large set of values  $(n, e)$  where each observed count  $n$  is a draw from a Poisson distribution with its own unknown mean  $\mu = \lambda e$ . Our goal is to evaluate the posterior distribution of each value of  $\lambda = \mu/e = E[R]$ . This statistical

framework, where there are a great many similar estimation problems that can be formulated within a common model, is a natural application of *empirical Bayes* methodology, sometimes also called *hierarchical modeling*. See the books [7] or [13] for extensive expositions of the methodology. DuMouchel et al [9] and [8] introduced the application of empirical Bayes estimation to data mining of large sparse frequency tables in the domains of natural language processing and screening of adverse drug reaction reports, respectively. The latter paper is accompanied by the discussion of two FDA scientists, Drs. Ana Szarfman and Robert O'Neill, who describe the adoption and successful application of the methodology at the FDA. Those two papers considered the analysis of large sparse two-dimensional tables, whereas in this paper we extend the methodology to  $J$  dimensions, where  $J$  is the item set size being considered. This Section provides an overview of the rationale and properties of the empirical Bayes method applied to the market basket problem, while a more formal development is deferred to the Appendix.

Empirical Bayes estimation enhances the simple use of the separate values of  $R = n/e$  as the estimates for each separate  $\lambda$  by adding the prior information that the many values of  $\lambda$  are connected in that they can be treated as having arisen from a common super-population of  $\lambda$ -values. The method assumes that the set of unknown  $\lambda$ s are distributed according to the continuous parametric density function of a specific form  $\pi(\lambda|\theta)$ . The parameter vector  $\theta$  is estimated from the data, the feature that distinguishes empirical Bayes methods from full Bayes methods. The estimation of  $\theta$  involves fitting the distribution of all the values of  $R$ , and once  $\theta$  has been estimated, the prior distribution of every  $\lambda$  is assumed to be  $\pi(\lambda|\hat{\theta})$ . Since this allows us to improve over the simple use of each  $R$  as estimates of the corresponding  $\lambda$ s, we say that the values of  $R$  "borrow strength" from each other to improve every estimate. To summarize, the empirical Bayes algorithm involves the following series of assumptions and steps:

1. A collection of pairs  $(n, e)$ , where each  $n > n^*$ , the minimum support.
2. Assume that conditional on known  $\lambda$ , each  $n$  is a Poisson random variable with mean  $\lambda e$ .
3. Assume that for some fixed  $\theta$ , the  $\lambda$ s are distributed according to a family of prior distributions,  $\pi(\lambda|\theta)$ . As discussed in the Appendix, we use a mixture of two gamma distributions as a convenient five-parameter family.
4. Compute the unconditional distribution of each  $n$  as  $f(n) = \int \text{Poi}(n|\lambda e) \pi(\lambda|\theta) d\lambda$ , where  $\text{Poi}(\cdot)$  is the Poisson density function.
5. Use the product of all expressions  $f(n)$  in Step 4 to compute the maximum likelihood estimate,  $\hat{\theta}$ .
6. For each  $(n, e)$  use Bayes rule to compute the posterior distribution of  $\lambda$  as  $\text{Poi}(n|\lambda e) \pi(\lambda|\hat{\theta}) / f(n)$ .
7. For each  $\lambda$ , use the results of Step 6 to obtain the posterior mean of  $\log(\lambda)$  and exponentiate it to obtain  $\Lambda$ , the empirical Bayes geometric mean of  $\lambda$ . Also obtain  $\lambda_{.05}$ , the

5<sup>th</sup> percentile of the posterior distribution of  $\lambda$ , for use as a Bayesian lower confidence limit.

Although the above procedure may seem complicated, as discussed in the Appendix, the amount of computation involved is small compared to the computation required just to tabulate the item set frequencies  $n$ . The resulting estimates  $\Lambda$  and lower confidence limits  $\lambda_{.05}$  are much more reliable than  $R$  when  $n$  and  $e$  are not large, and they will be very close to  $R$  when  $n$  and  $e$  are large. [8, Tables 4 and 6] compared three methods of ranking collections of pairs  $(n, e)$  (in the adverse drug reaction domain): based on the value of  $R$ , based on the value of  $\Lambda$ , and based on the significance level of the hypothesis test that  $n \sim \text{Poisson}(e)$ . The highest-ranking pairs according to  $R$  had  $n = 1$  or 2 but tiny  $e$ , such as  $(n = 1, e = .0003, R = 3300, \Lambda = 2.4)$ . The highest-ranking pairs according to statistical significance had huge  $n$ , but not very large values of  $R$ , such as  $(n = 7500, e = 500, R = 15, \Lambda = 15)$ . The highest-ranking pairs according to the empirical Bayes geometric mean were a compromise between these extremes, such as  $(n = 300, e = .2, R = 1500, \Lambda = 1200)$ . Ranking item sets according to their significance levels of an hypothesis test of independence tends to give too much weight to highest frequency item sets that may have only moderately high interest measures  $\lambda$ . We believe that in most domains,  $(n = 300, e = 0.2)$  is more “interesting” than is  $(n = 7500, e = 500)$ .

Sorting item sets of each size by the largest values of  $\Lambda$  or  $\lambda_{.05}$  is one recommended way to prioritize the examination of the modeling results. This assumes that the ratio parameter  $\lambda$  is of primary interest, and in our experience that assumption is reasonable when the main objective is to uncover causal relationships among the items. On the other hand, in many business applications the discovery value of an association may be roughly proportional to the number of times that the corresponding item set occurs in excess of that predicted by independence. In such a situation, we recommend sorting the item sets by the variable EXCESS, where

$$\text{EXCESS} = e \times (\lambda_{.05} - 1). \quad (1)$$

EXCESS is interpreted as a conservative (95% lower confidence limit) estimate of the number of times that the item set occurred in excess of the level of pure noise.

#### 4. MULTI-ITEM ASSOCIATION VERSUS PAIRWISE ASSOCIATION

For pairs of items, the interpretation of association, or deviation from independence, is straightforward, since there is only one degree of freedom for dependence and the single correlation between the two items is a familiar concept. For larger item sets the situation is more complicated. Contributing to the mutual dependence of  $J$  items, there are  $2^J - J - 1$  degrees of freedom, which can be partitioned into  $J(J-1)/2$  pairwise correlations and the remaining  $2^J - J(J+1)/2 - 1$  degrees of freedom for higher-order association. For triples, there are 3 pairwise associations and 1 degree of freedom for 3-factor interaction. For quadruples, there are 6 pairwise associations and 5 degrees of freedom for 3- and 4-factor interactions. When an analyst finds that the frequency of a triple or quadruple is much greater than independence predicts, what actually has been found? Suppose a triple ABC is unusually frequent. Is that just because AB and/or

AC and/or BC are unusually frequent, or is there something special about the triple that all three occur frequently in transactions? If the former, then the mining of triples hasn't really found anything that couldn't have been deduced from the results of examining pairs. If the latter, further investigation may uncover an important new insight into the domain. For example, in the medical domain, suppose in a database of patient adverse drug reactions, A and B are two drugs, and C is the occurrence of kidney failure. In case 1, A and B may act independently upon the kidney, but since A and B are sometimes prescribed together there may be many occurrences of ABC. In case 2, A and B may have no effect on the kidney if taken alone, but when taken together a drug interaction occurs that often leads to kidney failure. It would be valuable to be able to recognize case 2 situations automatically. In general, it would be quite valuable to automatically pick out the multi-item associations that cannot be explained by the pairwise associations in the item set. After all, there may be tens of thousands of supposedly interesting triples and quadruples to investigate after performing a market basket analysis, and it would be nice to know which of them can be discarded as “nothing new” once you have completed the analysis of pairs.

We will address this problem using the standard statistical theory of log-linear models as treated in textbooks like [4] or [5]. Consider the  $2^J$  frequency table determined by the joint distribution of any  $J > 2$  columns of  $X$ . The counts in the table are denoted  $m_{j_1 j_2 \dots j_J}$ , where each  $j_1, \dots, j_J$  is either 0 or 1. The  $J$  variables are said to obey the “all-two-factor” model if the expected frequencies of the array of counts are of the form

$$E[m_{j_1 j_2 \dots j_J}] = ab_1^{j_1} b_2^{j_2} \dots b_J^{j_J} c_{12}^{j_1 j_2} c_{13}^{j_1 j_3} c_{23}^{j_2 j_3} \dots c_{J-1, J}^{j_{J-1} j_J} \quad (2)$$

The model parameters are  $a, b_1, \dots, b_J, c_{12}, \dots, c_{J-1, J}$ . Thus, for the item set defined by these  $J$  items, the all-two-factor model predicts that  $E[m_{00\dots 0}] = a$  transactions contain none of the  $J$  items, and that  $E[m_{11\dots 1}] = ab_1 \dots b_J c_{12} \dots c_{J-1, J}$  transactions contain the complete item set. Log-linear model theory states that the all-two-factor model is the canonical model for determining a joint distribution from all the two-way marginal distributions without adding any more complex information. For example, the sufficient statistics for fitting the all-two-factor model are the  $J \times (J-1)/2$   $2 \times 2$  tables of counts from the two-way marginals. The fitted values (2) are easily computed using the iterative proportional fitting algorithm [4], which is guaranteed to converge. For any item set of size  $J > 2$ , whose frequency has been computed as  $n = m_{11\dots 1} \geq n^*$ , the frequencies of smaller included item sets will be already counted and available to supply the statistics necessary to fit the all-two-factor model and therefore its value of  $E[m_{11\dots 1}]$ . This leads to the following definitions for each item set of size greater than 2:

$$e_{\text{All2F}} = E[m_{11\dots 1}] = \text{predicted count of all-two-factor model based on all two-way distributions} \quad (3)$$

$$\text{EXCESS2} = \Lambda \times e - e_{\text{All2F}} \quad (4)$$

EXCESS2 is an estimate of the number of transactions containing the item set over and above those that can be explained by the pairwise associations of the items in the item set. Although EXCESS2 could be negative, we hesitate interpreting such item sets, focusing mainly on large values of



EXCESS2 as possibly indicating complex relationships involving more than pairwise association among the items of the item set.

Note that in computing (3), rather than use the raw (and possibly unstable) two-way counts to fit the all-two-factor model, we use the shrinkage estimates of two-way counts based on substituting  $\mu_{jk} = \Lambda_{jk} \times e_{jk}$  for the raw two-way count  $n_{jk}$ , for every pair of items  $(j, k)$  contained in the item set. The four elements of the  $2 \times 2$  table used are thus  $\mu_{jk}$ ,  $n_j - \mu_{jk}$ ,  $n_k - \mu_{jk}$ , and  $N - n_j - n_k + \mu_{jk}$ .

## 5. EXAMPLE: INTERNATIONAL CALLING BEHAVIOR

In this section we apply the ideas to an example drawn from telecommunications. We do not pursue a complete analysis of these data. Rather we highlight those aspects of the analysis that differentiates our work from that reported in the literature.

In support of security operations, AT&T maintains a calling signature for all accounts that make international calls. This signature contains features of calling behavior such as estimated calling rate (e.g., calls/week), call time (e.g., time-of-day and day-of-week), call duration, and call terminations (e.g., countries called) that serve as a baseline for detecting anomalous activity. We focus on the call terminations component of the signature that consists of a list of the top-5 recently and frequently called countries for each account. This list evolves through time using a probabilistic bumping algorithm, so that while the list itself has a fixed length, the labels and associated frequencies of those labels change as calls are made. The complete item set consists of the 228 countries where AT&T terminates international calls. We generate "market baskets" by taking monthly snapshots of the top-5 country list for each account for three consecutive months. Note that on average, our accounts call only two different countries per month so that for most accounts there are fewer than 5 countries (items) in their monthly "market basket".

There are about 7 million accounts that regularly make international calls. These accounts are stratified along several dimensions:

- account type: business/ residence/ casual(unknown)
- line type: multi-line/ single-line
- usage category: low/ medium/ high

These account descriptors define 18 strata corresponding to the unique combinations of their levels. The number of accounts in each stratum varies from a low of a few hundred to a high of 2 million. It is desirable to stratify the data since the communication behavior is known to be different across these variables. In searching for interesting country combinations, we would not want to highlight country combinations that were interesting solely due to the individual countries' popularity in the same strata.

Our data set consists of over 20 million transactions (7M/month  $\times$  3 months). We chose the minimum support size of  $n^* = 4$ , and by construction, we only examined item sets up to size  $J = 5$ . Each of the three monthly analyses required approximately 52 minutes of CPU time on a Pentium 750 MHz processor. Almost all of this time was spent generating the collection of item sets satisfying the minimum support requirement (including about 16,000 pairs, 120,000 triples, 150,000 4-tuples and 29,000 5-

tuples). About 3 of the 52 CPU minutes were devoted to the statistical computations, namely computing the maximum likelihood estimates of  $\theta$  and the statistics associated with the (empirical Bayes) posterior distributions, including the interestingness measures  $\Lambda$ ,  $\lambda_{.05}$ , EXCESS, and EXCESS2.

**Minimum support size.** A common rule of thumb for specifying the minimum support size is to use 0.1% of the number of transactions. Since each month has approximately 7 million transactions, this would limit the analysis to item sets with frequencies of 7,000 or greater. Applying this rule to item sets of size one (i.e., single countries) would eliminate over half (120 of 228) of the countries from the analysis. For applications in fraud detection and targeted marketing, this is clearly far too large.

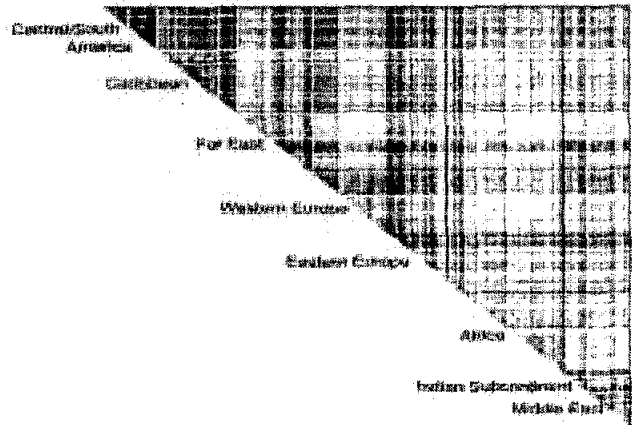


Figure 1. Shading of Country pairs based on raw counts  $n$ .

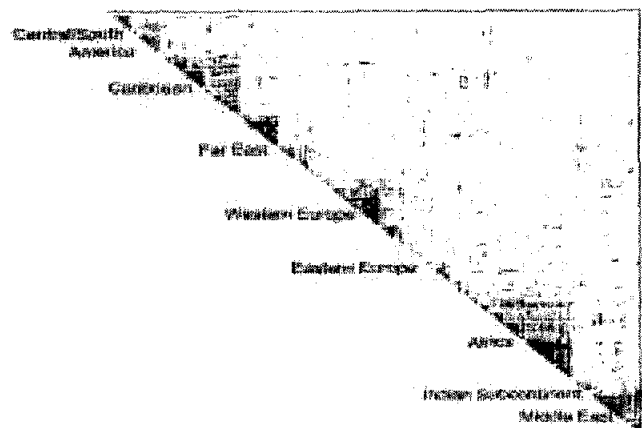


Figure 2. Shading of Country pairs based on empirical Bayes lower limit  $\lambda_{.05}$

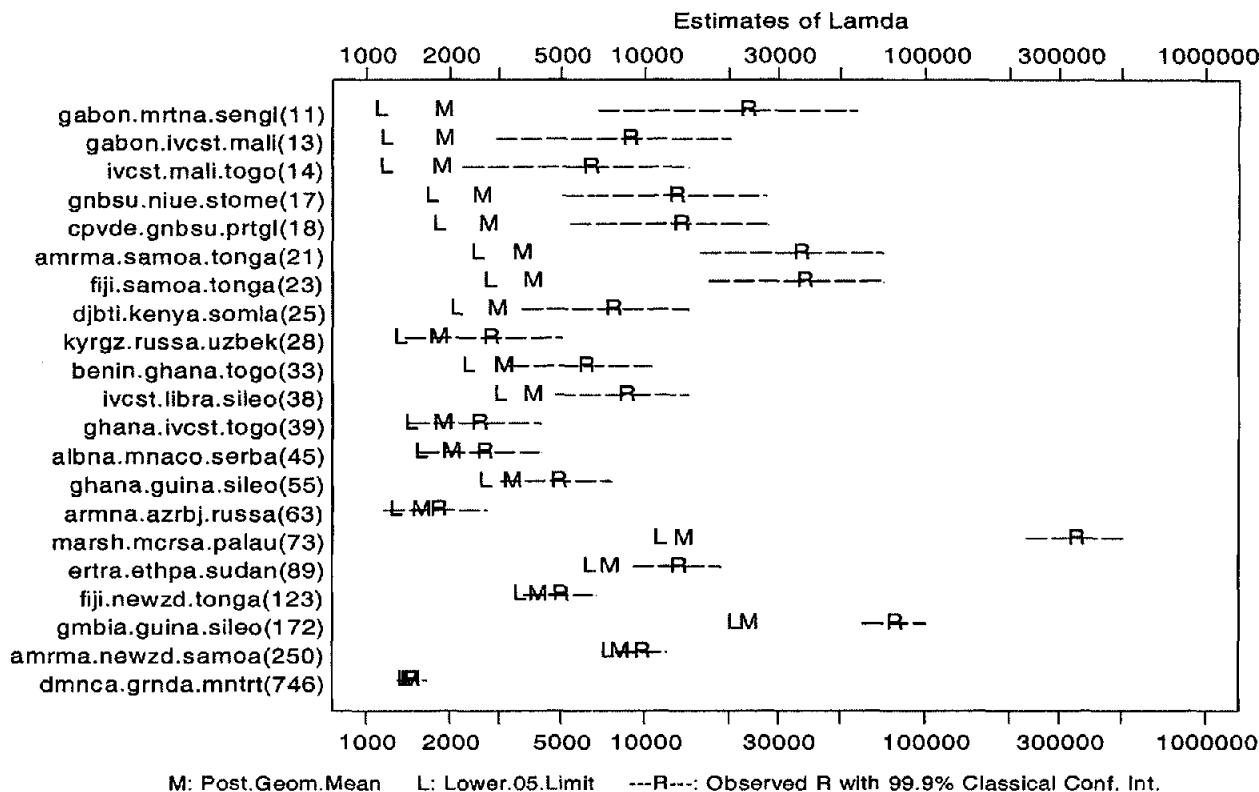
Figure 1 displays the support for item sets of size 2 (i.e., country pairs). The layout of the countries along the rows and columns roughly corresponds to geography. It was determined by a cluster analysis using an inter-country distance matrix based on an average  $\Lambda$  over the three months. We employ a gray scale to denote country pairs satisfying minimum support at four levels:

- |                                |                               |
|--------------------------------|-------------------------------|
| white: $n \leq 10$             | light gray: $10 < n \leq 100$ |
| dark gray: $100 < n \leq 1000$ | black: $1000 < n$             |

Only 1,633 of the 25,878 ( $=228 \times 227/2$ ) possible country pairs have support  $n > 1000$ . The situation gets more extreme as the size of the item sets increases. Lowering the minimum support requirement to  $n^* = 4$  leads to 66% of the country pairs being eligible for the subsequent analysis.

**Shrinkage.** Our proposed analysis focuses on estimates of the ratios of observed to baseline cell frequencies. Figure 2 displays the conservative estimate  $\lambda_{.05}$  averaged over the three successive months. The gray scale corresponds to increasing

$\lambda_{.05}$ , with lightest gray corresponding to  $2 \leq \lambda_{.05} < 4$  and the darkest to  $\lambda_{.05} \geq 16$ . White locations have either  $n < 4$  or  $\lambda_{.05} < 2$ . Just over 8% of the country pairs are shaded; this means that in these cells, our lower confidence limit has at least twice as many accounts calling this pair of countries than we would expect under the assumption that countries are called independently of each other. An interesting aspect of Figure 2 is that  $\lambda_{.05}$  is greatest for country pairs in the same part of the world — this is intuitively reasonable.



**Figure 3. Shrinkage for some month 2 triples having very large  $\lambda_{.05}$  and a wide range of  $n$  (in parentheses)**

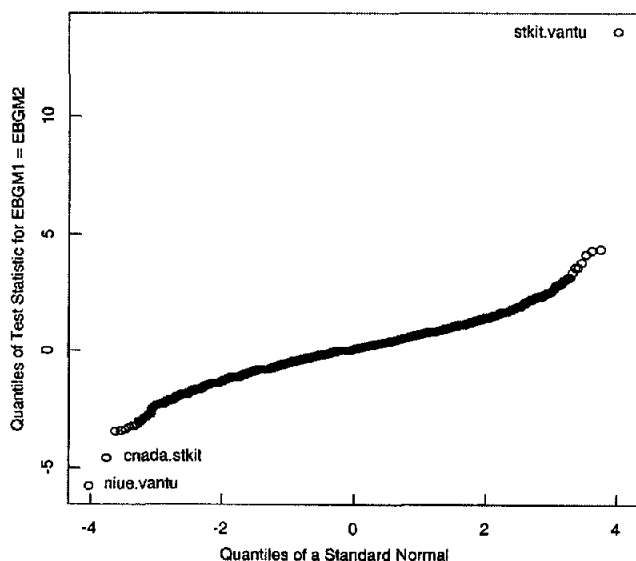
Figure 3 displays a comparison of the raw value of  $R = n/e$  with the empirical Bayes estimates  $\lambda_{.05}$  and  $\Lambda$  for some of the month 2 triples that have very large  $\lambda_{.05}$ . The triple of countries and the observed  $n$  of the triple are listed on the vertical axis. For each triple, the graph shows a classical 99.9% confidence interval for  $\lambda$ , based on the assumption that  $n$  is distributed  $\text{Poisson}(\lambda e)$ , as well as the two shrinkage estimators. The triples are ordered in increasing  $n$  from top to bottom of the graph, making it easy to see that for small  $n$  the shrinkage is more drastic than for larger  $n$ . One exception to this rule is the triple “marsh.mcrsa.palau” having ( $n = 73$ ,  $e = 0.0002$ ), which exhibits drastic shrinkage even though  $n$  is fairly large. This happens because the Bayesian estimate tends to shrink large  $R$  when  $e$  is very small, even if  $n$  is moderately large.

One way to ascertain the significance of individual measures when looking at a large collection is to use a quantile-quantile plot. This technique is standard in the statistics community as it allows one to visually calibrate a collection of measures that are all drawn from a common distribution. The intuition is that any

sample drawn from a common distribution will have a smallest and a largest value — the question is “How big is big?”. The quantile-quantile plot attempts to answer this question by displaying the quantiles (e.g., ordered values) of the observed measure against the quantiles of the assumed distribution of the measure. If the assumed distribution is correct, apart from random variation, the configuration of points should follow a straight line. Systematic departures from linearity indicate that the assumed distribution is not correct — common anomalies being systematic curvature or outliers at the extremes.

Figure 4 is an example of a “standard normal quantile” plot, so-called because the assumed distribution is a normal (i.e., Gaussian) distribution with mean zero and unit standard deviation. In it we display the ordered values of the standardized change in the  $\Lambda$  measure (also referred to as “EBGM”) from month one to month two. (Because our methods yield a posterior variance for every estimated  $\log \lambda$ , we can compute approximately standard normal test statistics for comparing the ratios of corresponding  $\lambda$ s from one month to the next.) Mild

systematic curvature suggests that the distribution of this measure is not Gaussian. The more relevant feature for this discussion however is the suggestion of a few outliers (labeled in the figure). We can provide a partial interpretation for the two extreme country pairs (St Kitts, Vanuatu) and (Niue, Vanuatu). All three countries are islands that host "adult entertainment" telecommunication services. The providers of these services earn commissions from the state-owned phone company for the traffic that they attract. In the case of (Niue, Vanuatu) there appears to have been a drop-off in the relative rate of accounts calling both of these countries in month two. The opposite is true for (St Kitts, Vanuatu) where there appears to be a surge in the relative rate of accounts that are calling both of these geographically dispersed islands. Indeed, callers are often referred to other "hot lines" and callers unaware of international dialing codes have little idea on where the calls terminate — until they receive their bill. This analysis provided us with our first indication that the adult entertainment business model is making a significant entry into the Caribbean area.



**Figure 4.** QQ plot for test statistics comparing each pair's  $A$  from month 1 to month 2.

*Multi-item versus pairwise association.* The interpretation of "interesting" large item sets can be confusing since it is often unclear whether the item set is interesting because it contains all the items, or if it is interesting because it consists of interesting subsets of items. Our models allow these alternatives to be compared by fitting all-two-factor log-linear models and comparing the fits to the shrinkage estimates.

Figure 5 displays the different measures of excess frequency of 4-tuples that result depending on whether or not we adjust for pairwise association. The figure displays the excess frequency for item sets of size four depending on whether comparison is to a complete independence model (abscissa = EXCESS) or the all-two-factor model (ordinate = EXCESS2). Figure 5 includes all 7,718 4-tuples in month 1 where both EXCESS and EXCESS2 are nonnegative. Several of the points in Figure 5

have been identified with their country names and the observed  $n$ . We comment on a few of them:

*(Austria, Germany, Mexico, UK):* This 4-tuple is unusual in that the independence model suggests that this item set is not so interesting (EXCESS almost 0), but when the pairwise interactions are taken into account, the interestingness of the 4-tuple is increased (EXCESS2 > 600). Since some of the pairwise associations are negative, the all-two-factor model predicts that this 4-tuple would occur somewhat less than independence predicts.

*(Australia, Canada, India, UK):* This 4-tuple has about 3600 more accounts calling them than independence predicts, but when pairwise interactions are taken into account, this 4-tuple has just about the frequency that we would expect.

*(China, Hong Kong, Mexico, Taiwan):* The 1585 accounts calling these four countries are far in excess of what either the independence model or the all-two-factor model predicts.

## 6. DISCUSSION

We introduced three variations on the market basket problem that are drawn from statistical considerations. First we introduced stratification of transactions by features that are known to correlate with item sets. Second we dramatically reduced the recommended minimal support size for item sets by introducing an empirical Bayes model that effectively takes into account variation associated with small frequencies. And finally we built on earlier work that considers interestingness measures that assess departures of observed frequencies from baseline frequencies. For item sets of size two, the independence model provides a natural baseline, but we argue that in larger item sets interpretations are sometimes aided by comparisons with fitted frequencies from other models, especially the log-linear model specifying all two-factor interactions. In every case, our methods provide for a reliable ranking of item sets by "interestingness." Given an interesting item set, our methods do not explicitly choose a predictive association rule. We prefer to let a domain expert make judgments about potential cause and effect or directional association. For example, our discovery that adult entertainment lines had spread to St. Kitts in the Caribbean was easy to verify once the outlier in Figure 4 was presented to someone familiar with the situation in Vanuatu.

We deliberately did not focus on algorithmic issues. In our experience the computing time to enumerate item sets and compute interestingness measures is negligible compared to the hours of analysis time spent perusing the voluminous output. We have developed and continue to develop visualization techniques to aid this phase of the problem. The figures in this paper are examples of some of the static displays that we feel are valuable. Yet the real power of visualization lies in interactive approaches, especially in high dimensions (i.e., large item sets). We are currently exploring such techniques and hope to report on this research at KDD2002.

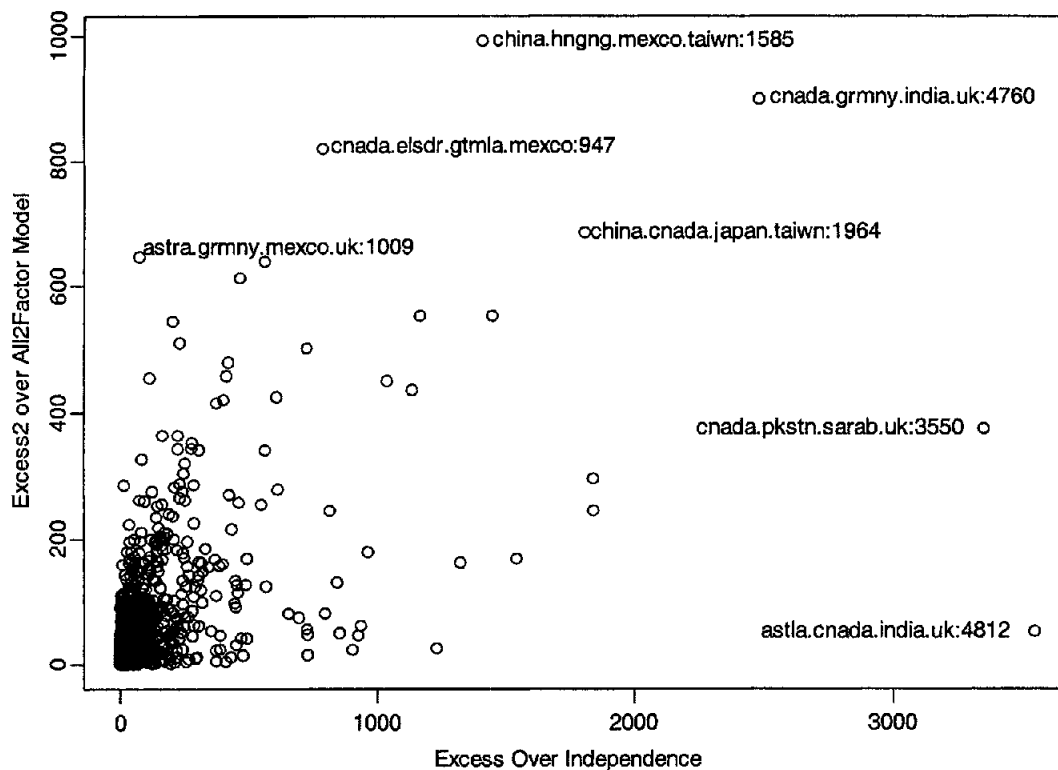


Figure 5. Excess vs. Excess2 for all 4-tuples where both are nonnegative (observed  $n$  listed after “:”).

## 7. REFERENCES

- [1] Aggarwal CC, Yu PS (1998) A new framework for item set generation. *Proc. of ACM-PODS Symposium on Principles of Database Systems*, Seattle, WA, pp. 18-24.
- [2] Agrawal R, Imilienski T, Swami A (1993) Mining association rules between sets of items in large databases. *Proc. ACM SIGMOD Intl. Conf. On Mgmt. of Data*, pp. 207-216.
- [3] Agrawal R, Srikant S (1994) Fast algorithms for mining association rules. In *Proc. 20<sup>th</sup> VLDB Conf, Santiago, Chile*.
- [4] Agresti A (1990) *Categorical Data Analysis*. New York: John Wiley.
- [5] Bishop YMM, Fienberg SE, Holland PW (1975) *Discrete Multivariate Analysis* Cambridge, MA: MIT Press.
- [6] Brin S, Motwani R, Ullman JD, Tsur S (1997) Dynamic item set counting and implication rules for market basket data. *Proc. ACM SIGMOD 1997 Intl. Conf. on Mgmt. of Data*, pp. 255-264.
- [7] Bryck A, Raudenbush S (1992) *Hierarchical Linear Models*. Newbury Park, CA: Sage Publications.
- [8] DuMouchel W (1999) Bayesian data mining in large frequency tables, with an application to the FDA Spontaneous Reporting System (with discussion), *The American Statistician*, 53:177-202.
- [9] DuMouchel W, Friedman C, Hripesak G, Johnson S, Clayton P (1996) Two applications of statistical modeling to natural language processing. *AI and Statistics V*, ch. 39, edited by D. Fisher and H. Lenz, Springer-Verlag.
- [10] DuMouchel W, Volinsky C, Johnson T, Cortes C, Pregibon D (1999) Squashing flat files flatter, *Proc. KDD 1999*, ACM Press, San Diego, CA, p. 6-15.
- [11] Johnson N, Kotz S (1969) *Discrete Distributions*. Houghton Mifflin, now distributed by New York: John Wiley.
- [12] Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* 22: 719-748.
- [13] O'Hagan A (1994) *Kendall's Advanced Theory of Statistics, vol. 2, Bayesian Inference*. New York: Halstead Press (John Wiley).
- [14] Silverstein C, Brin S, Motwani R (1998) Beyond market baskets: generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery* 2: 39-68.
- [15] Simpson EH (1951) The interpretation of interaction in contingency tables. *J. Royal Statistical Soc.*, B13:238-241.

## 8. APPENDIX: EMPIRICAL BAYES SHRINKAGE ESTIMATES

This development follows that of DuMouchel [8]. Rather than treat the separate values of  $\lambda$  as unrelated constants, assume that each  $\lambda$  is a random variable drawn from a common prior



distribution. This distribution is assumed be a mixture of two gamma distributions. The density function of a gamma distribution, having mean =  $\alpha/\beta$  and variance =  $\alpha/\beta^2$ , is

$$g(\lambda; \alpha, \beta) = \beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda} / \Gamma(\alpha)$$

The prior probability density of  $\lambda$  is assumed to be

$$\pi(\lambda; \alpha_1, \beta_1, \alpha_2, \beta_2, p) = p g(\lambda; \alpha_1, \beta_1) + (1-p) g(\lambda; \alpha_2, \beta_2) \quad (A1)$$

Therefore the prior mean of  $\lambda$  under this mixture model is  $p\alpha_1/\beta_1 + (1-p)\alpha_2/\beta_2$ , and its prior variance is  $p(1-p)(\alpha_1/\beta_1 - \alpha_2/\beta_2)^2 + p\alpha_1/\beta_1^2 + (1-p)\alpha_2/\beta_2^2$ . The exact choice of prior distribution for  $\lambda$  is not so important as that it have several free parameters so that the distribution of  $\lambda$  can be fit using observed data. The density in (1) has five free parameters:  $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, p)$ . The distributions for pairs, triples, quadruples, etc. have separate values of  $\theta$ . The family of gamma distributions are often used to model populations of Poisson rates because of the conjugate relationship between the Poisson and gamma distributions ([7][11][13]). The calculations are simplified in two respects: first, the marginal distribution of each  $n$  is a mixture of negative binomial distributions; and second, the posterior distribution of each  $\lambda$  is a mixture of two gamma distributions with modified parameters. Assuming that  $\theta$  and  $e$  are known, then the distribution of  $n$  is

$$\text{Prob}(n) = p f(n; \alpha_1, \beta_1, e) + (1-p) f(n; \alpha_2, \beta_2, e), \quad (A2)$$

$$f(n; \alpha, \beta, e) = (1 + \beta/e)^{-n} (1 + e/\beta)^{-\alpha} \Gamma(\alpha + n) / \Gamma(\alpha) n!$$

Let  $Q_n$  be the posterior probability that  $\lambda$  came from the first component of the mixture, given  $n$ . From Bayes rule, the formula for  $Q_n$  is

$$Q_n = p f(n; \alpha_1, \beta_1, e) / [p f(n; \alpha_1, \beta_1, e) + (1-p) f(n; \alpha_2, \beta_2, e)] \quad (A3)$$

The posterior distribution of  $\lambda$ , after observing  $n$ , can be represented as

$$\lambda | n \sim \pi(\lambda; \alpha_1 + n, \beta_1 + e, \alpha_2 + n, \beta_2 + e, Q_n) \quad (A4)$$

where  $\pi(\cdot)$  is given by (A1). Using well-known properties of gamma distributions, the posterior expectations of  $\lambda$  and  $\log(\lambda)$  are given by

$$E[\lambda | n] = Q_n (\alpha_1 + n) / (\beta_1 + e) + (1 - Q_n) (\alpha_2 + n) / (\beta_2 + e) \quad (A5)$$

$$E[\log(\lambda) | n] = Q_n [\psi(\alpha_1 + n) - \log(\beta_1 + e)] + (1 - Q_n) [\psi(\alpha_2 + n) - \log(\beta_2 + e)] \quad (A6)$$

where  $\psi(x)$  is the digamma function, the derivative of  $\log(\Gamma(x))$ . The posterior geometric mean (denoted EBGm in [8] and denoted  $\Lambda$  in this paper), of  $\lambda$  is then the exponential of (A6):

$$\Lambda = \exp(Q_n [\psi(\alpha_1 + n) - \log(\beta_1 + e)] + (1 - Q_n) [\psi(\alpha_2 + n) - \log(\beta_2 + e)]) \quad (A7)$$

This provides a "best" point estimate of each ratio  $\lambda$ , in the sense of minimizing squared error loss in  $\log(\lambda)$ . On the other hand, if a conservative estimate of  $\lambda$  is desired, in which the risk of overestimating every  $\lambda$  is constrained to be just 5%, then we can compute the 5<sup>th</sup> percentile of the posterior distribution of each  $\lambda$ , denoted  $\lambda_{.05}$ , as the solution to the equation:

$$0.05 = \int_0^{\lambda_{.05}} \pi(\lambda; \alpha_1 + n, \beta_1 + e, \alpha_2 + n, \beta_2 + e, Q_n) d\lambda \quad (A8)$$

where  $\pi(\cdot)$  is given by (A1). The integral in (A8) is easily computed using standard computer approximations of the incomplete gamma function, while the solution  $\lambda_{.05}$  requires an iterative technique such as Newton's method.

If  $e$  and  $n$  are both large, then  $\Lambda$  and  $\lambda_{.05}$  will both approach  $R = n/e$ , as can be seen from (A6), noting that for large arguments,  $\psi(x) \sim \log(x)$ , and  $\log(\alpha + n) - \log(\beta + e)$  will approach  $\log(n/e)$ . However, when  $e$  or  $n$  are not large, then the effect of using  $\Lambda$  is to "shrink"  $R$  toward smaller values, and the shrinkage is even greater for  $\lambda_{.05}$ . This is exactly the desired effect when sampling variation makes the true degree of association between the items in the item set uncertain.

To evaluate (A7) or (A8), estimates of  $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, p)$  are obtained by considering the marginal distribution of the set of all  $n$  for a given item set size, which is given in (A2). The negative binomial distributions  $f(\cdot)$  in (A2) are derived as a mixture of Poisson distributions, where the Poisson means have a gamma distribution [11, p. 125]. The likelihood function for  $\theta$  is the product of mixtures of two negative binomial densities:

$$L(\theta) = \prod \{p f(n; \alpha_1, \beta_1, e) + (1-p) f(n; \alpha_2, \beta_2, e)\} \quad (A9)$$

The maximum likelihood estimate of  $\theta$  is the vector that maximizes (A9), where the product in (A9) is over all possible item sets of the given size. The maximization involves an iterative search in the five dimensional parameter space, where each iteration involves computing  $\log[L(\theta)]$  and its first- and second-derivatives. Since  $\log[L(\theta)]$  would be the sum of an astronomical number of terms if we included all possible item sets (e.g., including item sets that were not observed in the database and have  $n = 0$ ), we use two modifications of (A9) to allow the procedure to scale up to large values of  $K$ , the number of unique items. First, we modify the likelihood to condition on  $n \geq n^*$ , where  $n^*$  is the minimum support that we wish to consider for item sets of the given size. This allows us to merely tabulate and include in the likelihood item sets having  $n \geq n^*$ , but, to properly accommodate this, we must replace  $f(n; \alpha, \beta, e)$  in (A9) by the conditional density of  $n$ , given that  $n \geq n^*$ , namely

$$f_*(n; \alpha, \beta, e, n^*) = f(n; \alpha, \beta, e) / [1 - \sum_{m=0}^{n^*-1} f(m; \alpha, \beta, e)] \quad (A10)$$

Even choosing  $n^* = 1$  saves an enormous amount of computation compared to enumerating all possible item sets and computing the corresponding values of  $e$  and  $f(\cdot)$ . As mentioned in the previous sections, we are focusing on applications where moderate values of  $n$  are of interest and so we take  $n^*$  to be at least 1 but as small as computational resources allow. If  $n^*$  is as large as several hundred, the noise in the counts for  $n \geq n^*$  will

typically be so little that the gain from computing  $\Lambda$  or  $\lambda_{.05}$  may be small.

In order to further reduce the computational burden of the iterative search to maximize  $L(\theta)$ , we further reduce the number of terms in (A9) by the technique of *data squashing* [10]. The collection of all item set descriptors  $(n, e)$ , where  $n \geq n^*$ , is replaced by a smaller set of  $M$  triples  $(n_i, e_i, w_i)$ ,  $i = 1, \dots, M$ . For each unique observed value of  $n$ , a set of values of  $(n_i = n, e_i, w_i)$  are chosen so that the distribution of the  $e_i$  weighted by the corresponding  $w_i$  mimics the distribution of the original values of  $e$  for that value of  $n$ . (A simple version of data squashing is to partition the values of  $e$  into bins for each unique value of  $n$  and replace the set of  $(n, e)$  in each bin by the single triple  $(n_i = n, e_i, w_i)$ , where  $e_i$  is the bin mean and  $w_i$  is the bin count of the  $i$ th bin.) This reduces the expression for  $L(\theta)$  to just  $M$  terms, namely:

$$L(\theta) = \prod_{i=1}^M [p f_*(n_i; \alpha_1, \beta_1, e_i, n^*) + (1-p) f_*(n_i; \alpha_2, \beta_2, e_i, n^*)]^{w_i}$$

In our experience where  $M$  is a few thousand, the maximization typically takes between 5 and 15 iterations from the starting point  $\theta_0 = (\alpha_1 = .1, \beta_1 = .1, \alpha_2 = 10, \beta_2 = 10, p = p_{max})$ , where  $p_{max}$  is the value from among (0.2, 0.5, 0.8) having the largest value of  $L(\theta_0)$ . [These starting values are used for item pairs,  $J = 2$ . Larger values of  $J$  tend to lead to more dispersed distributions of  $\lambda$ , so we divide the above starting values of the  $\alpha$ s and  $\beta$ s by  $3^{J-2}$ .] This experience also indicates that the computational effort of tabulating all the item sets having  $n > n^*$ , which uses an algorithm similar to the *Apriori* algorithm [2][3] involving one

pass over the transaction database for each item set size, is a much greater computational burden than the empirical Bayes calculations described in this Appendix and the log-linear model calculations discussed in Section 4. Thus there is no computation-related reason not to compute  $\Lambda$ ,  $\lambda_{.05}$ , EXCESS and EXCESS2 routinely.

Table 1 shows the estimates and standard errors of the hyper-parameters  $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, p)$  for each month and for the item set sizes 2–5. Standard errors are based on the second derivatives of the log likelihood function. It can be seen that the parameter values are quite consistent across databases, but that the values of  $\alpha$  and  $\beta$  tend to decrease as the size of the item set increases. This indicates that the distribution of  $\lambda$  becomes more and more dispersed as the dimension goes up, with more outlying item sets having very large frequencies compared to the baseline frequency based on the independence assumption. This also shows that the optimum amount of shrinkage varies depending on item set size.

We have not performed extensive goodness of fit analyses to the prior distribution family used here. One approach is to fit different strata separately and see if they lead to similar sets of hyper-parameters. If not, it is easy to combine results after separate fits to different subsets of the database or to different subsets of item sets. Classical hypothesis tests for goodness of fit to the observed distribution of  $(n, e)$  are possible but may not be useful, since massive data sets are bound to yield small "p-values" for rejecting almost any null hypothesis, even though the computed shrinkage estimates perform well as interestingness measures.

**Table 1. Hyper-parameter estimates and standard errors for the three months for dimensions 2 through 5.**

Month	Dim	Parameter	Alpha1	Beta1	p	Alpha2	Beta2	1-p
Oct	2	Estimate	0.1834	0.03935	0.2067	2.843	4.275	0.7933
		St.Error	0.01574	0.001915	0.005411	0.06065	0.1051	0.005411
Nov	2	Estimate	0.1845	0.03962	0.2068	2.857	4.219	0.7932
		St.Error	0.01594	0.001933	0.005448	0.06153	0.1047	0.005448
Dec	2	Estimate	0.1917	0.04174	0.2051	2.792	4.109	0.7949
		St.Error	0.01643	0.002039	0.005536	0.06052	0.1027	0.005536
Oct	3	Estimate	0.001012	0.004768	0.1806	1.061	2.075	0.8194
		St.Error	0.000193	0.000111	0.001951	0.0135	0.02612	0.001951
Nov	3	Estimate	0.001062	0.005027	0.1803	1.059	2.044	0.8197
		St.Error	0.0002021	0.000118	0.001978	0.0136	0.02595	0.001978
Dec	3	Estimate	0.001147	0.005523	0.1787	1.082	2.076	0.8213
		St.Error	0.000218	0.0001328	0.002002	0.01381	0.02616	0.002002
Oct	4	Estimate	0.000688	0.001037	0.1216	0.02398	0.8088	0.8784
		St.Error	0.000134	0.000026	0.001316	0.003049	0.007799	0.001316
Nov	4	Estimate	0.0008258	0.001493	0.127	0.03096	0.8202	0.873
		St.Error	0.0001558	0.000036	0.001396	0.003939	0.008709	0.001396
Dec	4	Estimate	0.0008782	0.001534	0.1218	0.02563	0.7966	0.8782
		St.Error	0.0001669	0.000040	0.001397	0.003259	0.007906	0.001397
Oct	5	Estimate	0.000258	0.000291	0.1164	0.02204	1.227	0.8836
		St.Error	0.000075	0.000026	0.002687	0.002775	0.02236	0.002687
Nov	5	Estimate	0.001313	0.000971	0.1307	0.02604	1.309	0.8693
		St.Error	0.0003172	0.000069	0.002942	0.003313	0.02427	0.002942
Dec	5	Estimate	0.00174	0.001474	0.1356	0.02851	1.322	0.8644
		St.Error	0.0004084	0.0001061	0.003155	0.003656	0.02539	0.003155