Modeling a Store's Product Space as a Social Network

Troy Raeder, Nitesh V. Chawla*
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, 46556 USA
{traeder, nchawla}@cse.nd.edu

Abstract

A market basket is a set of products that form a single retail transaction. This purchase data of products can shed important light on how product(s) might influence sales of other product(s). Departing from the standard approach of frequent itemset mining, we posit that purchase data can be modeled as a social network. One can then discover communities of products that are bought together, which can lead to expressive exploration and discovery of a larger influence zone of product(s). We develop a novel utility measure for communities of products and show, both financially and intuitively, that community detection provides a useful complement to association rules for market basket analysis. All our conclusions are validated on real store data.

1. Introduction

The field of *market basket analysis* studies the composition of the set of products that a customer purchases in a single shopping trip [15]. Associations can be discovered from this data to draw product item sets that tend to be sold together [1]. A famous example of this is the beer and diaper association. Any existing patterns or correlations among products are of great interest to retailers. A greater understanding of purchase behavior allows stores to better meet customers' needs and realize greater revenues.

In this paper, we take a different approach to the market basket problem, treating it as a search for *relationships* rather than associations. We build a social network out of the individual products in the store, (hereafter called a *product network*) and use *community detection* algorithms to isolate strong relationships. We study one structural characteristic of a product network, its degree distribution, to sug-

gest ways in which the composition of a product network differs from that of a traditional social network. Finally, we introduce a novel utility measure for communities of products and show that ranking communities by this measure can quickly isolate important relationships. We evaluate our methods both comparatively and financially, and conclude that community detection provides a valuable supplement to association rules. Our results are based on data from a University convenience store consisting of over 660,000 transactions involving over 2,200 products.

The rest of the paper is organized as follows. Section 2 discusses product networks and their properties. Section 3 introduces community detection and its application to market basket analysis, including our utility measure for communities. Section 4 presents our experimental setup and results, Section 5 discusses related work, and Section 6 closes with conclusions and recommendations.

2. Constructing a Network of Products

We begin our discussion by examining the properties of product networks and their similarities and differences with other types of social networks. To construct a network of products from a list of transactions, we follow an intuitive approach similar to that of several other authors [8, 10, 14]: each node in the network represents a product, and an edge appears between any two products that have been bought together in a transaction.

The networks discussed here and in the rest of the paper are based on transaction data collected from an on-campus convenience store at the University of Notre Dame during the calendar year 2006. The data contain complete transaction information, including date and time, products purchased, and total cost, for over 660,000 transactions involving 2,200 unique products. Due to privacy concerns, there is no way to associate transactions with individual people.

It has been well-established that real-world social networks often have *heavy-tailed* degree distributions, meaning that there are very few *hubs*, connected to hundreds or

thousands of others while the vast majority of nodes have very few neighbors. In our data, we find heavy-tailed behavior both locally and globally. Figure 1 shows the degree distribution of both the entire network and the distribution of edge weights around a single product. We see that both plots decay at least linearly on a log-log scale. Not only does the network have a heavy-tailed degree distribution, but that individual products do as well. This result suggests that the average product is bought infrequently with the majority of its neighbors, and frequently with only a few.

Figure 1 hints at the most difficult aspect of product networks in practice. They differ from more well-studied types of interaction networks for one simple reason: the presence of an edge does not necessarily imply a confirmed relationship between products. More well-studied networks, such as networks based on citations or phone calls, do not suffer this problem to nearly the same degree.

In citation networks, for example, two nodes linked together by an edge are necessarily related: if one paper cites another, there is a reason. A cell phone network will have a small number of incidental links, (wrong numbers, telemarketing, or random personal business), but most of the time, when one person calls another, it implies a connection between them. Product networks are different. Simply because a person buys paper towels and spaghetti sauce in the same transaction does not entail a common motivation for the two purchases. Worse, a person who buys several unrelated items in a single transaction will cause a clique to form between them, despite the absence of any true relationship.

As a result, product networks are very *dense*, with a large number of connections per node, but many of these edges are meaningless: representing spurious associations generated by chance. Our network contains 2,248 products and almost 250,000 edges between them. However, over 150,000 of these edges have a weight of one, meaning the two products were bought together only once in the entire year 2006, and over 235,000 have weight less than 10. These extremely low-weight edges are common and are unlikely to represent strong relationships.

In order to remove some of the noisy edges created by coincidental purchases and improve the quality of our subsequent analysis, we establish a minimum threshold σ, such that an edge exists between two products only if they have been bought together at least σ times. This is analogous to choosing a minimum support threshold for association rules. Note that, in the pruned network, the weight of an edge (p_1, p_2) remains the number of transactions in which p_1 and p_2 appear together.

Having described the construction of a product network and studied some of its properties, we now turn our attention to the analysis of the product space. Since the primary focus of market basket analysis is the discovery of relationships between products, we need to find groups of products whose structure or position within the network reveals useful information about the store itself.

Many real-world interaction networks naturally contain *communities*: groups of nodes that are more strongly connected to each other than they are to the rest of the network. Often, these communities have an easily-interpretable significance. In a cell phone network [16], for example, communities may represent families or circles of friends. Conversely, in a network of web pages [9] they may represent sites devoted to a common interest or theme. Community detection has been applied successfully in a numerous fields of science, ranging from social network analysis [16] to biology [2] and molecular physics [11]. It seems logical to expect that communities of products, since they are mutually strongly-connected, would be of particular interest. Therefore, the remainder of the paper will focus on the problem of *community detection* in product networks, and show how communities of products can be used to gain insight in to the behavior of customers in a store.

3. Discovering Communities of Products

Community detection is the process of finding strong communities in a network. The problem is usually addressed as follows: given a graph G, partition it into a series of disjoint subgraphs $\mathcal{G} = \{G_1 ... , G_n\}$ maximizing an objective function $f(\mathcal{G})$. The number of communities n is generally not known beforehand, but determined by the algorithm. Many community detection algorithms [3, 6, 12] attempt to optimize a quantity known as *modularity* [13]. The modularity Q of a set of communities is defined as: $Q = \sum_i (e_{ii} - a_i^2)$ where e_{ii} is the fraction of edges that join vertices in community i to other vertices in community i and a_i is the fraction of edge endpoints that lie in community i. Modularity measures the difference between the number of in-community edges in a given set of communities and the expected number of in-community edges in a random network with the same degree distribution.

This notion is very intuitive. If a set of communities has a large fraction of its edges falling within communities, (and therefore a relatively small fraction falling between communities), then that particular community decomposition probably represents a strong community structure.

The application to market basket analysis is clear: isolating tightly-connected communities within the network of products will allow us to identify strong relationships among the products and, therefore meaningful correlations in customer purchase behavior. Furthermore, because communities can be arbitrarily large, they should be able to represent these relationships much more expressively and with less redundancy than ordinary association rules.
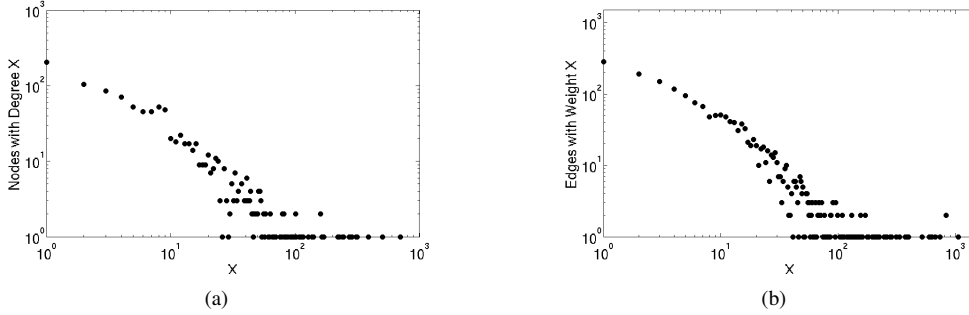
Figure 1. Degree distribution for (a) the entire network and (b) the neighbors of a single product.

Measuring the Utility of Communities Before we present our results, we quantify the utility of a community. Specifically, we wish to answer the question: *given a set of communities in a product network, which are most useful to a human analyst?*

Intuitively, the *utility* of a community can be determined by two opposing forces: *information*, and *information density*. A useful community will be large enough to provide a substantial insight into customer behavior, but small enough to be human-interpretable. To this end, we propose the following quantitative definitions. Define the *information* present in a community to be the sum, over all the edges in the community, of the *confidence* of the relationship indicated by the edge. The confidence of the relationship $A \to B$ is the observed conditional probability that B is purchased given that A is purchased.

$$I(G_i) = \sum_{(p_1, p_2) \in E_i} P(p_1 | p_2) \qquad (1)$$

We could have chosen, in lieu of confidence, a number of measures for the strength of an edge. The choice of confidence is convenient for two reasons. First, it is bounded. An unbounded measure, which can take values up to infinity, may assign an unreasonably high value to a community containing a single interesting relationship. Second, it is *null invariant* [17], meaning that its measure of the relationship between A and B is unaffected by transactions containing neither A nor B. To see why null invariance is important, consider two seasonal products that are sold only one month of the year. Even if these products are bought together 100% of the time, a measure that is not null-invariant (such as support) will likely see the relationship as weak because, for most of the year they are not bought at all.

Next, we define the *information density* $D(G_i)$ of community i as the information per node in G_i: $D(G_i) = \frac{I(G_i)}{|V_i|}$. Finally, we define the overall utility of community i as the harmonic mean of the above-defined quantities:

$$U(G_i) = \frac{2I(G_i)D(G_i)}{I(G_i) + D(G_i)}. \qquad (2)$$

Substituting the definitions of $I(G_i)$ and $D(G_i)$ into Equation 2 yields: $U(G_i) = D(G_i)\frac{|V_i|}{|V_i|+1}$. Thus, our measure prefers dense communities but given two communities of roughly equal density, it favors the larger one. This matches the intuition given earlier.

Because the computation in Equation 1 depends on the actual number of edges present in the community, our utility measure depends somewhat on the method of graph construction. In other words, if we allow an edge between any two products that are bought together, the computation will be different than if we restrict edges to products bought together at least 100 times. The end result of this is that our utility measure is *not* comparable across different network constructions. We do not consider this to be a significant issue because it is designed to help a human analyst assess one set of communities.

While our utility measure is designed for product networks, we believe that the tradeoff between size and density is very general and that, in principle, Equation 2 could be applied to other domains. In an email network, for example, if one defines *information* as the frequency of email correspondence between members of the community over some time period, an analog of Equation 2 follows naturally.

4. Experimental Evaluation

In order to demonstrate the effectiveness of our proposed methods, we present results from our 2006 data. We built a product network in the manner described above, setting minimum support at 0.01%. We present communities discovered with the algorithm of Blondel et al. [3], which is one of the more scalable algorithms available, and rank them using the measure defined in Equation 2. Though we use only one algorithm here, our studies have shown that differences across algorithms are largely insignificant.

(a) Chips and salsa.

(b) Eggs and baking products.
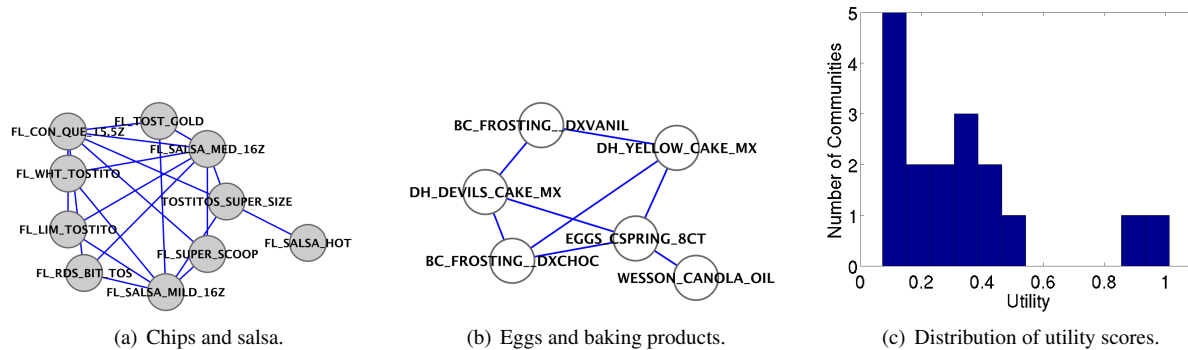
(c) Distribution of utility scores.

Figure 2. The first two communities in our data, ranked by the measure given in Equation 2, and the distribution of utility scores across all communities.

Overall, there were 17 communities discovered in the pruned network, ranging in size from two products to over 70. We evaluated each of these communities using the utility measure defined in Equation 2 and the results appear in Figure 2(c). The calculated utilities range from very near zero to slightly over 1. We see that a large number of communities have very low utility, with five communities falling in the first bin (below 0.14). At the other end of the spectrum, two communities rate substantially higher than the others (1.01 and 0.92 respectively). Highly-rated communities are generally well-connected with a clear purpose.

Figure 2(a) shows the most highest-rated community, a community of chips and salsa. The community is very densely connected, and it carries a very clear message: that people often buy chips and salsa together, and yet is small enough for a human to easily interpret. The community is nearly bipartite, with chips connecting only to salsa and salsa connecting only to chips. The one exception is a single edge between FL_SALSA_CON_QUE and FL_SALSA_MILD. From this community, it becomes clear that chips and salsa are *complementary* products, while the different types of chips (and respectively salsa) are *substitutes* for one another. The salsa con queso is an exception, because it is distinct from the other types available.

Figure 2(b) shows the second-ranked community, a collection of eggs and baking products. The structure of the community, with eggs as a hub in the center and the baking items the periphery, seems to imply that when people buy eggs in our store, they buy them for baking. Further investigation supports this initial hypothesis.

There were 541 distinct products bought with eggs at our store in the calendar year 2006, and in 18.5% of the cases, eggs were bought alone. However, at least one item among the six neighbors appears in over 39% of all transactions containing eggs, which is especially significant because most of the transactions in our store

are small. As a case study, we further quantify the impact of this particular community. Similar analysis can be applied to other communities, but space limitations preclude such analysis in this paper. Intuitively, cake mix is the most likely "causal" item in the group (it is unlikely, for example, that people buy frosting because they have a craving for eggs). Therefore, we calculate expected additional sales from each sale of cake mix as: $E(Sales) = P(Eggs|CakeMix) * Price(Eggs) + P(Frosting|CakeMix) * Price(Frosting)$ and find that the store can expect to generate $2.30 in additional sales from each cake mix sold. Therefore, the store stands to profit from any promotion that increases the sales of cake mix at a cost of less than $2.30 per transaction. Since cake mix itself costs $2.69, the expected additional revenue is 85.5% of the item's purchase price. This analysis is admittedly simple, but it demonstrates that communities can help identify profitable promotions in a store.

The third-and-fourth-ranked communities, shown in Figures 3(a) and 3(b) are communities of cereal and milk. The first of these shows a small container of milk as a hub surrounded by a series of cereals. In this case, the milk is small, at one pint, and many of the cereals are smaller individual-serving cereals. The second is composed of two nearly-disconnected subgraphs: a hub-and-spoke arrangement of larger milks and cereals and a clique of sodas. The disparate structures are each connected, by one edge, to a single product: plastic beverage cups.

These communities support several conclusions in addition to the notion that people buy cereal and milk together. First, there are separate relationships between cereal and milk at two levels: smaller sizes of milk correlate with smaller sizes of cereal, while larger milks relate to larger cereals. Second, the strong mutual correlation among sodas suggests that they are often purchased several at a time, while the disconnection among cereals indicates that people
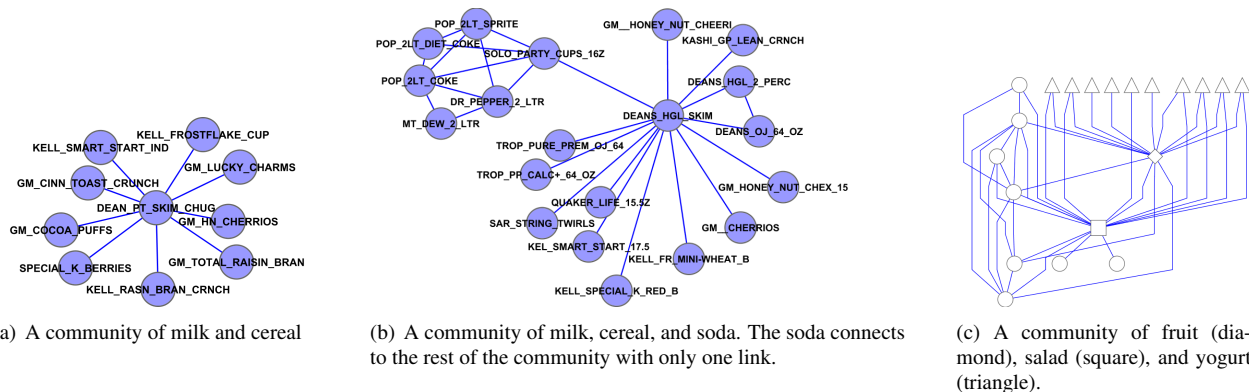
(a) A community of milk and cereal

(b) A community of milk, cereal, and soda. The soda connects to the rest of the community with only one link.

(c) A community of fruit (diamond), salad (square), and yogurt (triangle).

Figure 3. Three more communities.

buy them largely for personal use.

The final community of interest is shown in Figure 3(c). A community of fruit, salad, yogurt, and soft drinks, it is much less dense than the others and therefore, at number eight, is ranked much less favorably. However, it still contains useful insights. Figure 3(c) shows the single fruit product (diamond) connected to nine different yogurt products (triangles). The associations between fruit and any of the individual yogurt products are not strong (none is ranked better than 78th, in a list of 168 rules, by any of the interestingness measures in [17]), but in combination the association is quite powerful.

If all the different varieties of yogurt are combined, they become the most popular product purchased with fruit, and we find that 10% of all fruit sales (by dollar value) come in transactions that contain yogurt, and that 9.5% of all yogurt transactions contains some form of fruit. By contrast, if all varieties of coffee are combined, coffee (the runner-up) occurs in only 8% of fruit transactions, despite the fact that it is bought five times more frequently than yogurt overall. The fruit and yogurt association, then is a significant relationship whose significance is hidden by the number of yogurt products available.

The largest community, not shown for the sake of space, contains over 70 products. Composed of many of the store's most popular items, it is too large and dense to be easily interpreted. This fact, in conjunction with the communities mentioned above, suggests that community detection can play a useful supplementary role in market basket analysis. The highly-ranked communities discussed above provide a good deal of insight into the purchases of items as diverse as fruit, cereal, and frosting, but communities reveal very little with regard to the dense "core" of the network: popular products such as coffee, bagels, and water.

Therefore, we propose that community detection be used as a first exploratory step in the analysis process, where it will illuminate the relationships among important, but more peripheral, products. Then, the subsequent association rules analysis can focus more intently on products whose role is not clear within the community decomposition, perhaps with techniques like Association Rules Networks [4, 5].

5. Related Work

Association rules are a popular methodology for discovering frequent item sets in transactional data. However, Klemettinen et al [10] point out that association rules can be quickly limited due to high redundancy and many trivial rules, which can quickly overwhelm the user. As a result various researchers [7, 10, 17] have proposed different modifications and extensions to standard association rules.

One popular method is to mine *closed* or *maximal* itemsets. An itemset I is closed if no superset of I has the same support as I and maximal at $s\%$ support if no superset of I has at least $s\%$ support. The effectiveness of these methods in practice depends on the composition of the data. If a dataset supports several rules $A \rightarrow B, AC \rightarrow B, AD \rightarrow B, ...$ maximal itemset mining will prune the first of these rules but leave the others. If the first rule arises as a consequence of the others, then the pruning is useful. However, if the additional products C, D, etc. co-occur incidentally with the popular products A and B, then the remaining rules are the ones that are redundant. In practice, mining maximal or closed itemsets may do little to eliminate redundancy. In our data, we observe that the number of rules is reduced, but is not significantly reduced. At 0.01% support and 10% confidence, 155 of the 168 rules discovered are maximal. At 0.05% support, the numbers increase to 340 and 385 respectively. In both cases, all the itemsets are closed.

Interestingness measures [17] quantify the interestingness of individual rules, such that the most useful rules in a ruleset can be examined first. It has been shown, however, [17] that interestingness measures tend to rank rules inconsistently. As such, these metrics are most useful if

background knowledge suggests an appropriate measure.

Association Rules Networks (ARN) [4, 5] are directed hypergraphs built from sets of association rules. Generally speaking, given a set of association rules and a target product, an ARN shows the extent to which rules "flow into" the target product, mapping out both direct and indirect causes. The key difference between our work and ARN is that ARNs require a target product for drawing connections, while our work can be used for extensive exploration to discover relationships among a multitude of products without requiring a target product.

Hyperclique patterns [18] are very strongly-related sets of products, somewhat similar to communities. However, the criteria that define a hyperclique pattern are very strong in practice, and it is difficult to find hyperclique patterns of any substantial size. In our data, there are no hyperclique patterns of size greater than two, even at support as low as 0.005%. Therefore, hyperclique patterns and communities discover different kinds of relationships.

Clauset et al. [6] apply community detection to Amazon.com transaction data, but their treatment of the data is very basic. They do not explain any of the communities found, or address any practical issues, but merely state that the communities "make sense." Hao et al. [8] develop an application that uses networks to visualize association rules from e-commerce transaction data. Specifically, the application does a force-directed layout of the products in a network, and is capable of performing k-means clustering on the resulting visualization. Our approach is more general, in that community detection algorithms do not require users to specify the number of communities to find. Also, k-means can be sensitive to the initial locations of the cluster centers, which imposes an additional parameter on the process.

6. Conclusions

This paper studies the application of a social network model and community detection algorithms to the problem of *market basket analysis*: the search for meaningful relationships in purchase data. We posit that ties arising from random chance are an especially grave problem in product networks and suggest a thresholding approach for eliminating redundant ties. We find communities in our data that express complex relationships among products. These relationships illuminate the products' role in the store, but may also be potentially profitable. Additionally, we develop a novel measure of the utility of a community of products, and show that it favors communities that are large enough to be useful, but small enough to be interpretable. Finally, Our method of network construction requires only one parameter: a minimum support, and therefore is no more dependent upon parameters than traditional association rules.

Because community detection seems to be most effec-

tive at determining relationships between products outside the core of the network, we recommend the use of community detection as a first exploratory step in conjunction with association rule analysis methods.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in very large databases. In *Proc. 20th International Conference on VLDB*, pages 487–499, 1994.

[2] S. Asur, D. Ucar, and S. Parthasarathy. An ensemble framework for clustering protein-protein interaction networks. In *ISMB/ECCB*, pages 29–40, 2007.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks, 2008.

[4] S. Chawla, B. Arunasalam, and J. Davis. Mining open source software (oss) data using association rules network. *PAKDD*, pages 461–466, 2003.

[5] S. Chawla, J. Davis, and G. Pandey. On Local Pruning of Association Rules Using Directed Hypergraphs. *20th International Conference on Data Engeneering*, 2004.

[6] A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70(066111), 2004.

[7] K. Gouda and M. Zaki. Efficiently mining maximal frequent itemsets. In *Proceedings of ICDM*, pages 163–170. IEEE Computer Society, 2001.

[8] M. Hao, U. Dayal, M. Hsu, T. Sprenger, and M. Gross. Visualization of directed associations in e-commerce transaction data. *Proceedings of VisSym*, 1:185–192, 2001.

[9] J. Kleinberg and S. Lawrence. The structure of the web. *Science*, 294:1849–1850, 11 2001.

[10] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. Verkamo. Finding interesting rules from large sets of discovered association rules. *Proceedings of CIKM*, pages 401–407, 1994.

[11] C. Massen and J. Doye. Identifying communities within energy landscapes. *Physical Review E*, 71(4):46101, 2005.

[12] M. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):36104, 2006.

[13] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):26113, 2004.

[14] C. Palmer and C. Faloutsos. Electricity Based External Similarity of Categorical Attributes. *LNCS*, pages 486–500, 2003.

[15] G. Russell and A. Petersen. Analysis of cross category dependence in market basket selection. *Journal of Retailing*, 76(3):367–392, 2000.

[16] K. Steinhaeuser and N. Chawla. Community detection in a large-scale real world social network. In *LNCS*. Springer Verlag, 2008.

[17] P. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.

[18] H. Xiong, P. Tan, and V. Kumar. Hyperclique pattern discovery. *Data Mining and Knowledge Discovery*, 13(2):219–242, 2006.