Soc. Netw. Anal. Min. (2011) 1:97–113
DOI 10.1007/s13278-010-0003-7

ORIGINAL ARTICLE

Market basket analysis with networks

Troy Raeder · Nitesh V. Chawla

Received: 10 March 2010 / Accepted: 6 July 2010 / Published online: 28 August 2010

Abstract The field of *market basket analysis*, the search for meaningful associations in customer purchase data, is one of the oldest areas of data mining. The typical solution involves the mining and analysis of *association rules*, which take the form of statements such as "people who buy diapers are likely to buy beer". It is well-known, however, that typical transaction datasets can support hundreds or thousands of obvious association rules for each interesting rule, and filtering through the rules is a non-trivial task (Klemettinen et al. In: Proceedings of CIKM, pp 401–407, 1994). One may use an interestingness measure to quantify the usefulness of various rules, but there is no single agreed-upon measure and different measures can result in very different rankings of association rules. In this work, we take a different approach to mining transaction data. By modeling the data as a *product network*, we discover expressive communities (clusters) in the data, which can then be targeted for further analysis. We demonstrate that our network based approach can concisely isolate influence among products, mitigating the need to search through massive lists of association rules. We develop an interestingness measure for communities of products and show that it isolates useful, actionable communities. Finally, we build upon our experience with product networks to propose a comprehensive analysis strategy by combining both traditional and network-based techniques. This framework is capable of generating insights that are difficult to achieve with traditional analysis methods.

Keywords Market basket analysis · Community detection · Product network · Transaction data · Association rules

1 Introduction

The collection and study of retail transaction data, known as *market basket analysis*, has become increasingly prevalent in the past several years. Many supermarkets, for example, issue *loyalty cards* (Mauri 2003). While providing discounts to the customer, these cards allow the retailer to develop a better understanding of individuals' purchasing habits by associating customers with transactions. The uses of this information vary, but may include informing product placement decisions, designing personalized marketing campaigns, and determining the timing and extent of product promotions (Adomavicius and Tuzhilin 1999; Agrawal and Srikant 1994; Cho et al. 2002) among others.

Formally, the task of market basket analysis is to discover actionable knowledge in transaction databases. The problem can be understood as follows: a standard retail store sells a large set of products $\mathbf{P}$. Define a *transaction* $\mathbf{p} \subseteq \mathbf{P}$ as the set of products an individual customer buys in a single trip to the store. The store's *transaction database* $\mathbf{T} = \{\mathbf{p}\}$ is the set of all transactions the store has processed within a given time period. Ultimately, an effective analysis method should enable the retailer to draw clear, comprehensive conclusions from the data.

One popular tool for market basket analysis in practice is the mining of *association rules* (Agrawal and Srikant 1994). A set of association rules $R(\mathbf{T}, s, c)$ is defined by a

T. Raeder (✉) · N. V. Chawla
Department of Computer Science and Engineering,
Interdisciplinary Center for Network Science and Applications,
University of Notre Dame, Notre Dame, IN 46556, USA
e-mail: traeder@cse.nd.edu

N. V. Chawla
e-mail: nchawla@cse.nd.edu

transaction database **T**, a *minimum support* parameter s and a *minimum confidence* parameter c. Define A and B as arbitrary sets of products. Further, define **A** (analogously **B**) as the set of transactions containing every product in A (B). Formally, R is the set of all rules $A \rightarrow B$ such that:

1. $\frac{|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{T}|} \geq s$
2. $\frac{|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{A}|} \geq c$.

Association rules have found successful application in many diverse contexts and a number of algorithms have been developed to discover them efficiently (Agrawal and Srikant 1994; Brin et al. 1997b; Hipp et al. 2000; Zaki 1999), but they are not without limitations. The most prominent of these is sheer volume. Large transaction datasets tend to contain hundreds or thousands of rules at reasonable levels of support and confidence, and many of these may be redundant or obvious (Klemettinen et al. 1994). As a result, it is often difficult to isolate interesting relationships.

Two distinct classes of methods have evolved to address this problem. One class (Gouda and Zaki 2001; Klemettinen et al. 1994; Zaki 2000; Zaki and Hsiao 2002) attempts to eliminate any rules that may be redundant, while the other (DuMouchel and Pregibon 2001; McGarry 2005; Tan et al. 2004) aims to elevate rules that are especially interesting (by sorting on an objective measure). Unfortunately, the concepts of both interestingness and redundancy are somewhat subjective. As a result, (which we show in Sect. 2) these methods are of limited use in practice.

Ultimately, existing literature on market basket analysis has failed to provide conclusive answers to some of the field's most pressing questions. For example, there is no widely-accepted means of isolating representative or useful relationships in market basket datasets and no existing work of which we are aware has attempted to offer any manner of procedural guidance for analyzing such data. In other words, no work has addressed the question *Given a new market basket dataset, what method or methods should I apply in order to obtain effective insights?*

This work attempts to address these concerns and improve the power and clarity of market basket analysis by modeling transactional data as a network. We show that by detecting *communities* of products in this network, we can discover strong and expressive relationships among products including relationships that are difficult to discover with traditional association rules. We then build on our experience with product networks and with a number of different market-basket and graph-theoretic algorithms to propose a novel procedure for mining unseen market basket datasets. The network representation of transaction data allows for the use of a diverse array of algorithms previously unavailable to the association rule community. As a result, this procedure is the first comprehensive market basket analysis framework ever proposed in the literature. All of our developments and conclusions are verified on real transaction data, consisting of over 660,000 transactions across more than 2,200 items, from an on-campus convenience store at the University of Notre Dame.

The remainder of the paper is organized as follows: Sect. 2 explores the strengths and weaknesses of traditional association rules analysis on our transaction data. The results presented here motivate the rest of the paper and serve as an introduction to the data itself. Section 3 introduces the concept of product networks and presents some properties of our network. Section 4 describes our community detection approach to market basket analysis and presents the first known interestingness measure for communities of products. Section 5 develops a comprehensive and novel framework for market basket analysis, incorporating both techniques introduced in this paper and previously-developed network analysis methods. Finally, Sect. 6 acknowledges some related work not mentioned elsewhere in the paper and Sect. 7 concludes.

2 Association rules

A popular approach for analyzing market basket data is the discovery and interpretation of *association rules*. The association rules problem (Agrawal and Srikant 1994) is defined as follows:

Given a threshold s, called the *minimum support* and a threshold c, the *minimum confidence*, find all rules of the form $A \rightarrow B$, where A and B are sets of products, such that:

1. A and B appear together in at least $s\%$ of transactions.
2. B occurs in at least $c\%$ of the transactions in which A occurs.

Sets of products are typically called *itemsets*, itemsets of size k are called k-itemsets, and sets that meet the minimum support criterion are typically called *large* or *frequent* itemsets. An association rule is said to be *supported* in a transaction database if it meets both the minimum support and minimum confidence criteria.

Algorithms for efficiently enumerating association rules are well-known (Agrawal and Srikant 1994; Han and Pei 2000; Zaki et al. 1997) and are a popular tool for unsupervised data exploration. As they came into widespread use, researchers noticed that understanding the rules themselves was not a trivial matter. First, there is no obvious method for choosing appropriate support and confidence thresholds. If the thresholds are chosen too high, interesting associations may be missed. However, if they are chosen too low, the user may be inundated with thousands of weak rules that do not represent meaningful associations.

To illustrate the magnitude of this problem, and in particular the difficulty of isolating appropriate thresholds, we discovered association rules in our own data at varying levels of support and confidence. Figure 1a shows the number of association rules discovered at 10% confidence as support ranges from 0.005 to 1%. The number of rules is negligible above 0.1% support but increases very rapidly below 0.05%. Figure 1b shows a similar result, this time holding support steady at 0.01% and varying confidence from 5 to 100%. The increase appears substantially less drastic but this is largely due to a number of redundant multi-item associations with exceptionally high confidence. Note that from 10 to 5%, the number of rules more than doubles. Taken together, Fig. 1a, b show that association rules can be incredibly sensitive to the choice of support and confidence parameters.

A second practical issue is that transaction databases often contain hundreds or thousands of association rules at reasonable levels of support and confidence, and many of those rules are either redundant or simply obvious (Klemettinen et al. 1994).

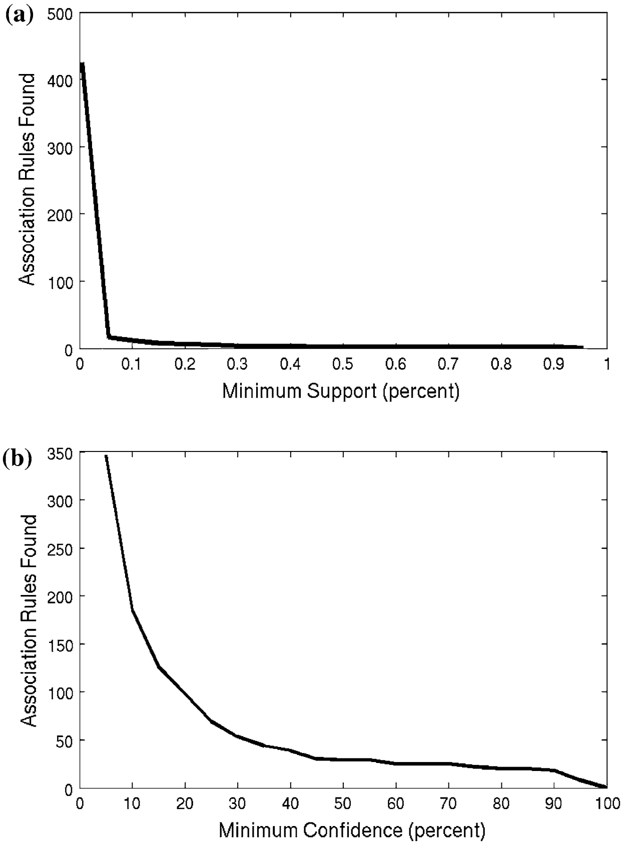


Fig. 1 Number of associations discovered at varying levels of confidence and support: **a** 10% confidence, varying support; **b** 0.01% support, varying confidence

A number of different techniques have been developed to address this issue. The first is the mining of *maximal* (Gouda and Zaki 2001) or *closed* (Zaki 2000; Zaki and Hsiao 2002) itemsets. An itemset I is closed if no superset of I has the same support as I and I is 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. Furthermore, the number of pruned rules may be very small compared to the number of rules remaining.

As an example, our data supports 168 rules at 0.01% support and 10% confidence. Of these rules, 155 are maximal. Decreasing support to 0.005%, the numbers increase to 385 and 340 respectively. In both cases, all the itemsets are closed. Also, of the original 168 rules, 38 take the form {CREAM_CHEESE, X} → BAGEL or {BAGEL, X} → CREAM_CHEESE. Within these rules, all are closed and only two, (BAGEL → CREAM_CHEESE and CREAM_CHEESE → BAGEL) are not maximal. This result suggests that, in addition to pruning very few rules, maximal itemset mining, in our case, prunes incorrectly. Those 36 rules involving bagel and cream cheese can be very effectively explained by the very strong relationship between cream cheese and bagel.[1]

These findings may seem to be in conflict with prior research on closed and maximal itemsets. For example, in Zaki (2000), the author claims that the mining of closed itemsets can reduce the number of association rules found in a dataset by as much as a factor of 3,000. Those experiments, however, were conducted on generic machine learning datasets rather than market basket datasets. Furthermore, the results were obtained by mining association rules with multi-item consequents, which is rarely done in practice because it is known to produce redundant rules. We believe based on our results that maximal and closed itemset mining are of limited use for practical market basket analysis.

A second approach to combat the explosion of uninteresting rules is to calculate additional *interestingness* measures (Tan et al. 2004) on the rules. These measures can

[1] *A matter of notation:* Throughout the paper, as we discuss insights from our data, it will be necessary to mention a number of specific products sold in the store. Whenever we do so, we will denote them in ALL_CAPS to distinguish specific products from concepts or classes of items. Classes of items are typed in normal text. Thus, throughout the paper, WATER_DASANI_20_OZ refers to a specific type of water, whereas "water" refers to a general class of products.

then be used to either rank the rules by importance (and present a sorted list to the user) or as an additional pruning criterion. How exactly interestingness is determined varies by measure, but many existing measures take the approach that interestingness is "deviation from independence". For example, one of the simpler such measures, the *lift* of a rule $A \rightarrow B$ is defined as:

$$L(A \rightarrow B) = \frac{P(AB)}{P(A)P(B)} \tag{1}$$

where $P(X)$ is the proportion of transactions in which X occurs. Note that if purchases of A and B are perfectly independent, the lift $L(A \rightarrow B) = 1$. If A and B appear together more often than we would expect under independence, the lift is greater than 1, and otherwise it is less than one.

This notion of interestingness is intuitively reasonable, but there are dozens of such measures defined in the literature (Brijs et al. 2003; DuMouchel and Pregibon 2001; Tan et al. 2004) and it has been shown that they tend to rank rules very differently (Tan et al. 2004). Therefore, it is not obvious a priori which measure, if any, will elevate the desired rules to the top, or at what level of interestingness the useful rules will end.

To study this phenomenon in our own data, we found association rules at 0.01% support and 10% confidence and ranked them according to each measure given by Tan et al. (2004). Table 1 shows information about the top ten rules by average rank. Four of these rules are ranked best by at least one measure, and one ranks as badly as 129. Even the relationship between BAGEL and CREAM_CHEESE, which is the strongest in the data (support almost 1%, confidence 93%) is ranked 128th by one measure. This variability implies that interestingness measures are useful mainly when experience or background knowledge is available to assist in the selection of an appropriate measure.

An alternative approach to searching through large sets of rules is to impose a pruning criterion that preserves only the strongest relationships in the data. Hyperclique patterns (Xiong et al. 2006) discover tightly-knit groups of items, potentially at a much lower level of support than is feasible with association rules. A hyperclique pattern P at support s and *h-confidence* c is a set of items $P = \{P_1, P_2, \ldots, P_n\}$ such that for each association rule $P_i \rightarrow P_1 \ldots P_{i-1}, P_{i+1} \ldots P_n$, the support of the rule is at least s and the confidence of the rule is at least c. The advantage of hyperclique patterns is that they are able to discover relevant patterns without an explosion of the rule-space, as might be with using vanilla association rules. 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 market basket data. For our data, there are no hyperclique patterns of size greater than two, even at support as low as 0.005%. Therefore hyperclique patterns, while effective at discovering certain strong relationships, are hardly a sufficient analysis technique on their own.

Association rules networks (Chawla et al. 2003, 2004; Pandey et al. 2009) reduce the ruleset by focusing solely on rules related to a single product. More specifically, given a set of association rules R and a target product z, the association rules network ARN (R, z) is the unique directed hypergraph G satisfying the following properties:

1. Any hyperedge in G corresponds to a rule in R with a one-item consequent.
2. There is a hyperedge corresponding to a rule whose consequent is the target product z.
3. The target product z is reachable from every vertex v in G.
4. No vertex $v \neq z$ is reachable from z.

Table 1 High, low, and mean rank and standard deviation of ranks for the top 10 rules by average rank among the 21 interestingness measures in Tan et al. (2004)

Rule	High	Low	Mean	SD
CREAM_CHEESE → BAGEL	1	128	18.07	33.79
Cake Mix[a] → Frosting	3	65.5	21.85	19.61
VAULT_SODA → VAULT_ZERO	6	71	24.95	15.77
YORK_MINT_PATTIES, DIET_COKE_20_OZ → NEWSPAPER_CHICAGO_TR	2	96	28.05	25.67
NEWSPAPER_CHICAGO_TR, DIET_COKE_20_OZ → YORK_MINT_PATTIES	8	96	28.85	22.90
BAGEL → CREAM_CHEESE	1	129	31.37	36.32
CREAM_CHEESE, COFFEE_12_OZ → BAGEL	1	133	32.02	42.90
NYQUIL → DAYQUIL	1	118.5	33.35	33.92
VAULT_ZERO → VAULT_SODA	16	70	33.40	13.41
Frosting → Cake Mix	3	69	34.37	43.42

[a] Product names are: DH_YELLOW_CAKE_MX_18 and DH_FROSTING_DXCHOC

Generally speaking, an ARN shows the extent to which rules “flow into” the target product. The resulting network can show both direct and indirect associations of the target product *z*. However, association rules networks can be quite sensitive to the choice of target product, and there is no obvious proper choice. As a result, one must have some idea of the products he or she is interested in before association rules networks are applicable. We explore the integration of association rules networks into a broader strategy for market basket analysis in Sect. 5.3.

The above discussion suggests that no technique currently available in the literature sufficiently addresses the problem of finding meaningful relationships in large transaction databases. This deficiency motivates our discussion of network methods for market basket analysis, which is the subject of the next section. We do not claim to definitively solve the market basket problem. However, we will show that as a first exploratory step, our techniques can discover expressive relationships from which we can draw direct conclusions about the nature of customer behavior in a store.

3 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 (Hao et al. 2001; Klemettinen et al. 1994; Palmer and Faloutsos 2003): 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.

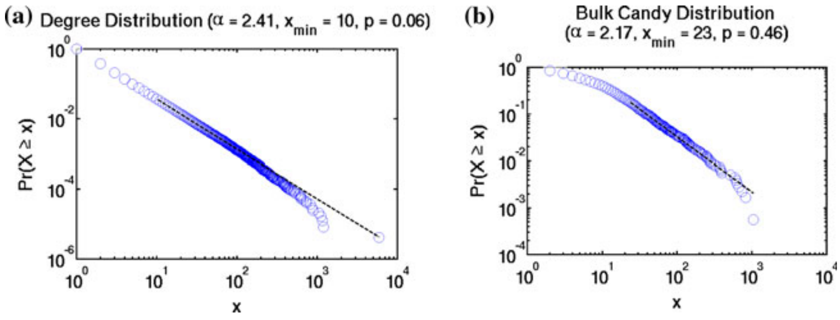
It has been well-established that real-world social networks often have *heavy-tailed* degree distributions, meaning that there are very few *hubs*, connected to many others while the vast majority of nodes have very few neighbors (Barabasi and Bonabeau 2003). In our data, we find heavy-tailed behavior both locally and globally. Figure 2 shows the degree distribution of the entire network and the distribution of edge weights around a single product. Each plot also contains best-fit power-law distributions calculated by the method of Clauset et al. (2007). The KS-test *p* values, given in the figures, show that the are both power laws at 0.05% confidence, although the degree distribution of the entire network is not nearly as strong of a fit. In any case, both distributions exhibit “heavy-tailed” behavior, in that the distributions are very heavily skewed toward small numbers but span many orders of magnitude. This result suggests that the average product is bought infrequently with the majority of its neighbors, and frequently with only a few.

Figure 2 hints at the most difficult aspect of product networks in practice. They differ from other types of interaction networks for one simple reason: the presence of an edge does not necessarily imply a confirmed relationship between products. Networks based on citations or phone calls, for example, do not suffer this problem to nearly the same degree.

In citation networks, 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 form a clique among 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

Fig. 2 Degree distribution for **a** the entire network and **b** the neighbors of a single product



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. One natural consequence of this density, many popular network statistics are unusually skewed. For example our product network has a 90% effective diameter of 4 and a full diameter of 5, much smaller than we would expect in a social network of the same size, and the average clustering coefficient is relatively high at 0.518.

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 the any remaining edge is unchanged.

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 (Steinhaeuser and Chawla 2008), for example, communities may represent families or circles of friends. Conversely, in a network of web pages (Kleinberg and Lawrence 2001) 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 (Steinhaeuser and Chawla 2008) to biology (Asur et al. 2007) and molecular physics (Massen and Doye 2005). 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.

4 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 $\mathcal{G} = \{G_1, \ldots, G_n\}$ 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 (Blondel et al. 2008; Clauset et al. 2004; Newman 2006) attempt to optimize a quantity known as *modularity* (Newman and Girvan 2004). The modularity Q of a set of communities is defined as:

$$Q = \sum_i (e_{ii} - a_i^2) \tag{2}$$

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.

4.1 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 $A \rightarrow B$ is the observed conditional probability that B is purchased given that A is purchased.

$$I(G_i) = \sum_{(p_1,p_2)\in E_i} P(p_1|p_2). \tag{3}$$

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* (Tan et al. 2004), 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:

$$D(G_i) = \frac{I(G_i)}{|V_i|}. \tag{4}$$

Finally, we define the overall utility of community i as the harmonic mean of the above-defined quantities:

$$U(G_i) = \frac{2I(G_i)D(G_i)}{I(G_i) + D(G_i)}. \tag{5}$$

Substituting the definitions of $I(G_i)$ and $D(G_i)$ into (5) 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 (3) 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, (5) 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 (5) follows naturally.

4.2 Results on real-world data

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 the support parameter $\sigma = 65$ (0.01% of all transactions). We present communities discovered with the algorithm of Blondel et al. (2008), which is one of the more scalable algorithms available, and rank them using the measure defined in (5). Though we use only one algorithm here, our studies have shown that differences across algorithms are largely insignificant.

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 (5) and the results appear in Fig. 3. 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 4a shows the most highest-rated community, consisting of different types 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 salsa con queso (`FL_SALSA_CON_QUE`) and medium salsa (`FL_SALSA_MED_16OZ`). 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 4b shows the second-ranked community, a collection of eggs and baking products. The structure of the community, with eggs (`EGGS_CSPRING_8CT`) 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

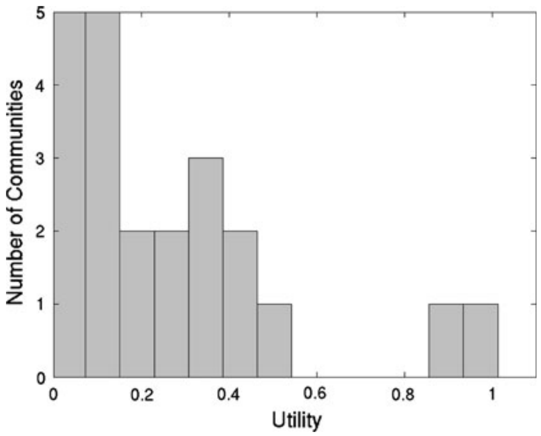


Fig. 3 The distribution of utility scores across all communities

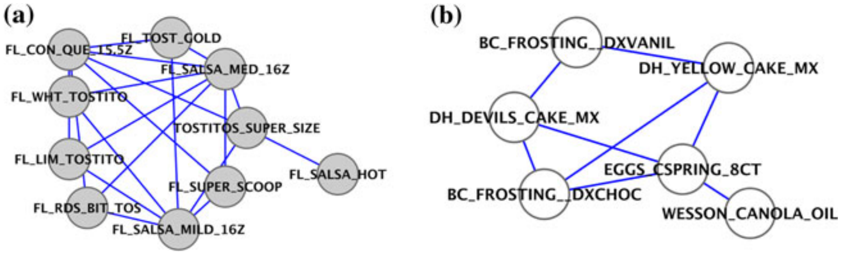


Fig. 4 The first two communities in our data, ranked by the measure given in (5). **a** Chips and salsa **b** Eggs and baking products

them for baking. Further investigation supports this initial hypothesis.

There were 541 distinct products bought with `EGGS_CSPRING_8CT` at our store in the calendar year 2006, and in 18.5% of the cases, they were bought alone. However, at least one item among the six neighbors appears in over 39% of all transactions containing `EGGS_CSPRING_8CT`, 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:

$$\begin{aligned} E(\text{Sales}) = {} & P(\text{Eggs}|\text{Cake Mix}) * \text{Price}(\text{Eggs}) \\ & + P(\text{Frosting}|\text{Cake Mix}) * \text{Price}(\text{Frosting}) \end{aligned}$$

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 Fig. 5a, 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 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 buy them largely for personal use.

The final community of interest is shown in Fig. 5c: a community containing fruit, salad, yogurt. 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 5c 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 Tan et al. (2004), 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, 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

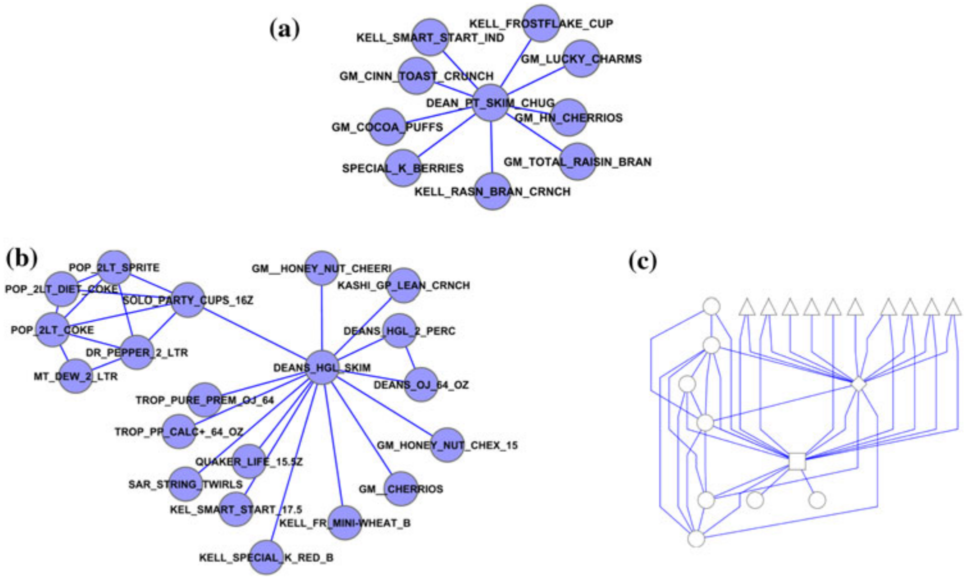


Fig. 5 Three more communities. **a** A community of milk and cereal. **b** A community of milk, cereal, and soda. The soda connects to the rest of the community with only one link. **c** A community of fruit (*diamond*), salad (*square*), and yogurt (*triangle*)

whose role is not clear within the community decomposition. The next section describes in greater detail our proposed framework for such an analysis.

5 Toward a comprehensive analysis strategy

A great deal of literature has been published on the subject of market basket analysis and survey papers about algorithms (Hipp et al. 2000; Zaki 1999), interestingness measures (McGarry 2005; Tan et al. 2004), and visualization techniques (Blanchard et al. 2003, Sect. 2) abound. In spite of all this effort, however, the community has made no substantive attempt to answer the following basic question: *Given a fresh, unseen market basket dataset what method or set of methods should be employed to obtain quick, actionable results?* There are several possible reasons for this. The first is a dearth of widely available transaction data, which we alluded to in the introduction. The second is a general lack of diversity in analysis techniques: maximal itemset mining, for example, is not different enough from traditional association rules such that the techniques can be *complementary*, with one strong where the other is weak. Finally, most studies that do consider real data are only conducted within a single domain (i.e. supermarkets or online retailers), and so the ability to draw overarching conclusions is limited.

Since we too are confined to a single dataset, we cannot address the third concern, but this section addresses the first and the second. In doing so, we call upon not only the techniques developed here in Sect. 3, but also a series of methods developed by other authors. To our knowledge, these methods [*association rules networks* (Chawla et al. 2003, 2004; Pandey et al. 2009) and *Center-Piece Subgraphs* (Tong and Faloutsos 2006)] have not been generally applied to market basket data, but in the course of our work we have found that they complement community detection nicely.

The rest of the section is organized as follows: Sect. 5.1 explores practical concerns regarding the use of association rules networks (introduced in Sect. 3), 5.2 introduces the Center-Piece Subgraph problem and studies its application in the domain of product networks, Sect. 5.3 ties together the discussion of this section and the prior one in order to propose a unified strategy for mining market basket data, and Sect. 5.4 briefly discusses strategies for parameter selection.

5.1 Association rules networks

Recall from Sect. 3 that an *association rules network* *ARN*(R, z) is a directed hypergraph representation of the ruleset R that mops out the direct and indirect associations of the target product z. The concerns we must address when applying association rules networks are (1) how do we choose an appropriate ruleset R? and (2) how do we choose an appropriate item z? The first question essentially boils down to the appropriate choice of support and confidence parameters, and we do not address it here. With regard to the second question, it is natural first to ask: is the choice of z important? Figure 6 shows two different association rules networks. In Fig. 6a, eggs are used as the target product, and in Fig. 6b, we use cake mix. Even though the two products chosen are related, we see that the resulting networks are quite different. While Fig. 6a shows a relationship between oil, eggs, cake mix and frosting, similar to

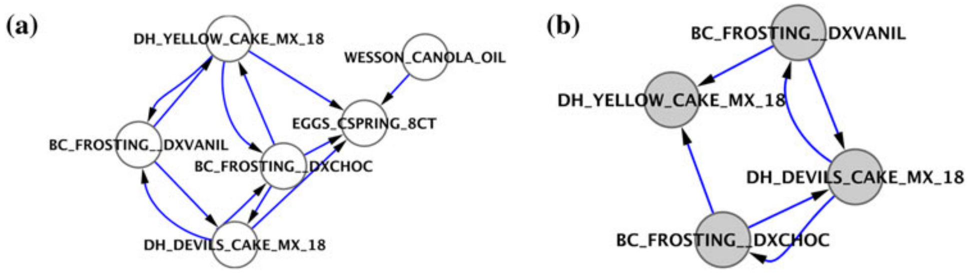


Fig. 6 Two association rules networks from the community of eggs. **a** Association rules network with $z =$ eggs (EGGS_CSPRING_8CT). **b** Association rules network with $z =$ cake mix (DH_YELLOW_CAKE_MX_18)

what was found with community detection, Fig. 6b contains only cake mix and frosting.

While Fig. 6 makes it clear that the target product z cannot be chosen arbitrarily, it does not shed any light on the process for making an appropriate choice. Figure 7 shows a separate association rules network flowing into BAGEL: one of the most popular items in the store. This network is large and expressive, and includes two of the relationships, fruit-yogurt and cereal-milk that we found with communities earlier (although not to the same detail). It provides an effective visualization of the relationships between some of the more central products in the store.

Many of the items that appear in the network, such as newspapers and donuts, are items that we would intuitively expect to sell well in the mornings. A cursory glance at the network suggests that coffee may drive food sales during the morning hours and bagels may drive drink sales. coffee does not connect to any other drinks, whereas BAGEL connects to drinks almost exclusively. Additionally, the network provides insight into the key relationships other core products: milk (with cereal), salad (with soup), and fruit (with salad and yogurt).

To understand why this bagel network is so much more informative than the cake mix network described above, we need to understand the ruleset on which the network is built. Recall that there were three rules containing BAGEL in the top 10 rules given in Table 1. As one would expect, the full ruleset contains substantially more. In fact, 47 of the 168 rules discovered contain BAGEL as either the antecedent or the consequent. This great diversity among BAGEL's "neighbors" in the network allows its ARN to span different segments of the product space.

Thus, it appears that an effective choice for z, when constructing an association rules network from transaction data, is to choose the item that appears in the most rules in the underlying ruleset R. One might consider instead the most popular product in the store, or the item which has been bought with the greatest number of other products. In our data, however, these strategies are less effective. BULK_CANDY, which is both the most frequently sold and bought with the most items, has only two products in its association rules network, and one popular type of water (WATER_DASANI_20_OZ), has none.

The reason for this is that association rules involving BULK_CANDY and WATER_DASANI_20_OZ, which are bought with a stunningly wide variety of items, do not meet the minimum confidence criterion that we have used throughout the paper. We contend, however, that relationships which do not meet the minimum confidence criterion may still be interesting. There are several potential causes of low confidence, but the most relevant in the case of water is *substitution*. There are many different types of water available in the store, and this variety erodes the confidence of certain relationships.

To illustrate the effect of substitution on rule confidence, assume n different products $F_1,\ldots,F_n$ are all substitutes for each other, meaning that they serve roughly the same function F. Furthermore, assume a product $\mathcal{P}$ correlates with items of the function F, such that the confidence of the association rule $F \rightarrow \mathcal{P}$ is c or

$$\frac{|F\mathcal{P}|}{|\mathcal{P}|} = c. \tag{6}$$

If the products $F_1,\ldots,F_n$ are all bought equally with $\mathcal{P}$, then for any F_i, the confidence of the rule $F_i \rightarrow \mathcal{P}$ is given by

$$\frac{\frac{|F\mathcal{P}|}{n}}{|\mathcal{P}|} = \frac{c}{n}. \tag{7}$$

Thus, the substitution erodes the confidence of the association $F_i \rightarrow \mathcal{P}$ even though the overarching association $F \rightarrow \mathcal{P}$ may be sufficiently interesting. It is also trivially true that substitution erodes the support of any relationship.

This parameter sensitivity is a problem inherent to every technique we have covered thus far. Association rules, ARNs, and the community detection framework we have defined will all systematically fail to find relationships that fall outside the specified support and confidence thresholds for any reason (substitution or otherwise). To address this issue, we turn to Center-piece subgraphs.

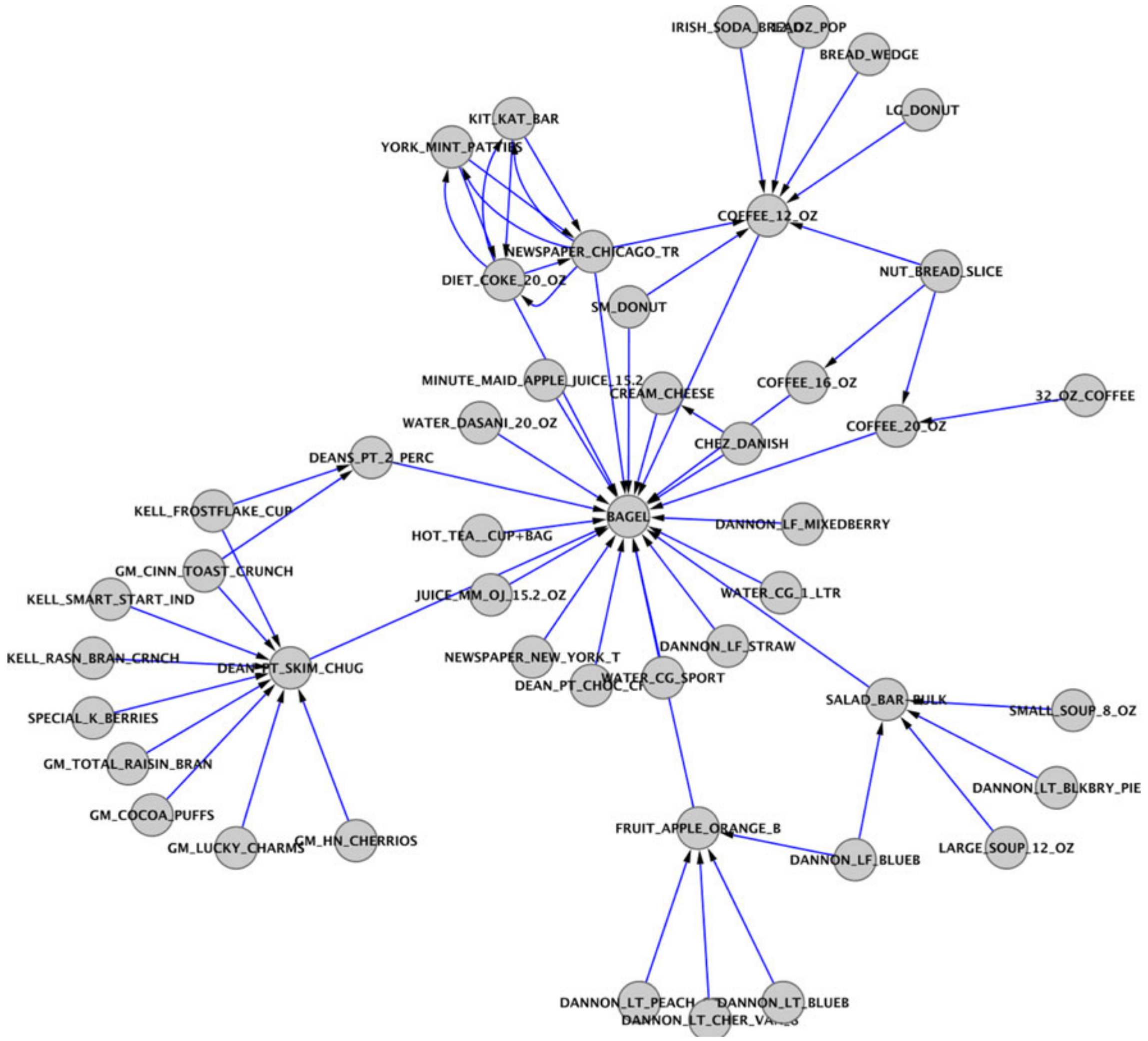


Fig. 7 Association rules network with z = BAGEL

5.2 Center-piece subgraphs

Center-piece subgraphs (CePS) (Tong and Faloutsos 2006), like association rules networks, describe the neighborhood of a node or set of nodes, but they differ considerably in how they define this neighborhood. The *Center-piece subgraph* $C_p(G, b, Q, k)$ is a subgraph $\mathcal{H}$ of the graph G that contains all *query* nodes in the set Q, contains at most b other nodes, and maximizes an objective function $g(\mathcal{H})$. The parameter k is called a soft_AND coefficient. In simple terms, k is the number of query nodes to which a node must be strongly related in order to be considered a candidate for the subgraph.

In other words, association rules networks define the "neighborhood" of the target product z as the set of set of products that are either direct or indirect causes of z within the ruleset R. A center-piece subgraph, by contrast, defines the neighborhood of the query nodes Q as the set of b products that are most closely related to the members of Q according to the objective function $g()$.

The benefit of center-piece subgraphs in the context of market basket analysis is that they allow us to trade scope for granularity. While community detection can find relationships in the product network with virtually no guidance, it requires a reasonable support threshold in order to isolate useful relationships. Similarly, association rules networks require the specification of a ruleset which is, by definition, constrained by a minimum support and confidence. Thus, in both cases, the *number* of products about which one can learn useful information is significantly constrained.

Center-piece subgraphs provide the opportunity to consider *all* products in an analysis, because the budget parameter b constrains the size of the sets that can be discovered. The cost of this added power is a tremendous decrease in scope. Whereas communities can discover relationships anywhere in the network, and an ARN may extend several levels out from the target product (recall the `BAGEL` ARN of Fig. 7), a center-piece subgraph is constrained to the set Q of query nodes and at most b other related products. As a result, the set of query nodes Q must be carefully defined in order for the resulting subgraph to be meaningful.

The remainder of the section discusses the objective function $g(\mathcal{H})$ maximized by CePS and outlines practical concerns regarding its application to market basket data. We conclude that, for the reasons stated above, center-piece subgraphs are primarily useful for either verification of hypotheses suggested by other techniques or for explaining unexpected results arrived at by other methods. For both of these applications, the set of query nodes Q will be very well-defined.

5.2.1 *Objective function definition*

Define a *Random Walk with Restart* (RWR) (Tong et al. 2006) on the graph G starting from a node $n \in V(G)$ as follows: At time t, a randomly-walking particle existing at node $n_t \in V(G)$ $(n_0 = n)$ transmits itself to one of the neighbors of n_t with a probability proportional to the weight of its edge with n_t. At any time, the particle has a fixed probability c of returning to node n.

From the normalized matrix of edge weights $\mathbf{W}$, one can calculate the probability $p_{i,j}^{(t)}$ that a randomly-walking particle starting at node i stands at j after exactly t steps. The limit as $t \rightarrow \infty$ of the $p_{i,j}^{(t)}$ is known as the *steady-state probability* that a particle starting at i will exist at node j. The vector of steady-state probabilities originating from node i, $\mathbf{p}_i$ can be calculated as Tong et al. (2006):

$$\mathbf{p}_i = c\mathbf{W}\mathbf{p}_i + (1 - c)\mathbf{e}_i. \tag{8}$$

where $\mathbf{e}_i$ is an indicator vector that is 1 in the i'th position and zero everywhere else. The matrix $\mathbf{W}$ is *normalized* in the sense that it is a transition matrix: i.e. $\mathbf{W}_{i,j}$ represents the probability that the randomly-walking particle will transition from i to j independent of the possibility of restart.

The RWR problem is very general and has been applied in a number of contexts. For example PageRank (Brin et al. 1998) now incorporates the notion of restart in its random-walk determination of page relevance to prevent assigning outrageous scores to dense communities of web pages. The CePS problem incorporates RWR into its goodness function as follows:

Define $r(i, j) = p_{i,j}$ to be the steady-state probability that a RWR starting at i exists at j. Further, define $r(Q, j, k)$ to be the steady-state probability that at least k RWRs originating from nodes in the query set Q simultaneously meet at node j. For "hard AND" queries, which are the type of query we will be most interested in, we can define the probability $r(Q, j)$ that random walkers from all query nodes meet at j as:

$$r(Q,j) = \prod_{i \in Q} r(i,j). \tag{9}$$

The objective function $g(\mathcal{H})$ for a subgraph $\mathcal{H}$ follows as:

$$g(\mathcal{H}) = \sum_{v \in V(\mathcal{H})} r(Q, v) \tag{10}$$

Tong and Faloutsos (2006) provide a fast algorithm for extracting subgraphs with high $g(\mathcal{H})$, and our experience shows that it scales to networks with thousands of nodes. In the next section we explore practical concerns regarding the application of CePS to market basket analysis and present results from our data.

5.2.2 *Center-piece subgraphs on market basket data*

Each technique we have discussed to this point has been limited by the need to specify a minimum support (and possibly minimum confidence) with which to discover relationships. As a result, strong relationships with low levels of support and substitution relationships with artificially low confidence are undiscovered.

Because center-piece subgraphs are constrained in size by the budget parameter b, it is unnecessary to further constrain them with minimum support and confidence parameters. As a result, they are the only technique we have discussed which is capable of discovering relationships between any and all products that make up the product space. The remainder of the section will show that this property makes center-piece subgraphs invaluable for the exploration of results obtained through other means. Specifically, they are effective for either verifying hypotheses suggested by other techniques or explaining relationships that do not, on the surface make sense.

Figure 8a shows a center-piece subgraph constructed from the full 2006 product network using a type of tortilla chips (`TOSTITOS_SUPER_SIZE`) as the query node and a budget $b = 10$. The network contains other chips and salsa, as our prior experience would lead us to expect, but also contains some items (`BULK_CANDY` and `BAGEL`) that are marginally related at best. We explored this phenomenon by constructing subgraphs of gradually increasing size in order to determine which items the algorithm considered more "important" with respect to the tortilla chips. In doing so, we found that the `BULK_CANDY` was added as the 6th member of the subgraph, before other products to which the chips have a stronger connection.

The reason for this is that `BULK_CANDY`, as a popular product, is bought with a tremendously large array of other products (recall the degree distribution of Sect. 3). To see

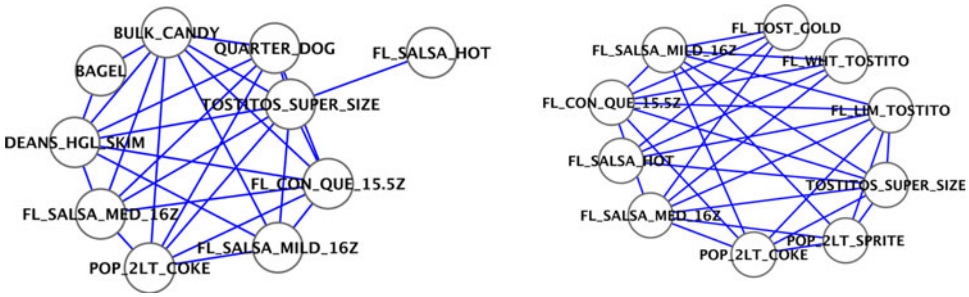


Fig. 8 Center-piece subgraphs with tortilla chips (TOSTITOS_SUPER_SIZE) as the query node. Edges are weighed by **a** support and **b** confidence

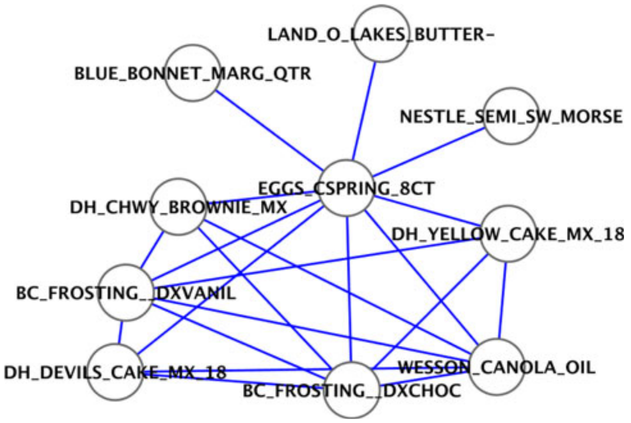


Fig. 9 A center-piece subgraph with eggs (EGGS_CSPRING_8CT) as the query node

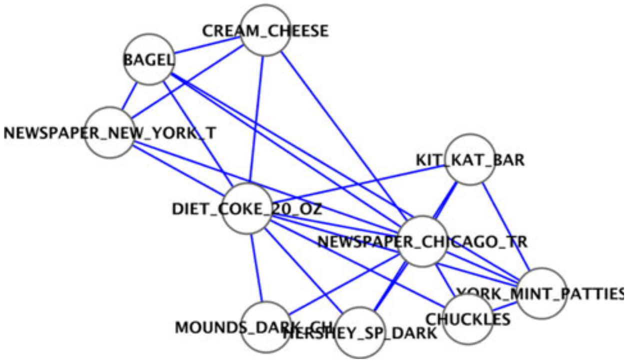


Fig. 10 A center-piece subgraph with Diet Coke (DIET_COKE_20_OZ), Newspaper (NEWSPAPER_CHICAGO_TR), and Peppermint Patties (YORK_MINT_PATTIES) as query nodes, to explain the association rule

why this causes problems for CePS, imagine a seldom-sold product p_j, appearing in five transactions, with which BULK_CANDY is bought once. By standard normalization, the transition probability *from* p_j *to* BULK_CANDY is at least 1/5, meaning that any random particle that reaches j is highly likely to reach BULK_CANDY. Combining this effect over hundreds of less popular products results in a very substantial steady-state probability for popular products.

To reduce the influence of such products, we weighted the edges by confidence instead of by absolute support. That is, the edge A–B is weighted with min($P(A|B)$, $P(B|A)$). There are two distinct advantages to using confidence in this instance. First, it forces all edge weights onto a uniform scale between zero and one. Second, it lessens the impact of coincidental purchases with popular products. In the example of the previous paragraph, the weight of the edge between p_j and BULK_CANDY is now $\approx \frac{1}{60{,}000}$ and after normalization it is likely that the transition probability from p_j to BULK_CANDY is much lower.

Figure 8b shows the impact of weighting edges by confidence. Now, instead of extraneous products like BAGEL and BULK_CANDY, we see sodas and other types of chips, which much more closely matches our intuition and corroborates the results found with other techniques.

Figure 9 shows a center-piece subgraph with eggs (EGGS_CSPRING_8CT) as the lone query node and a budget of 10. When we examined the community of eggs and cake mix in Sect. 3 we concluded that when customers bought eggs in our store, they bought them for baking. The subgraph in Fig. 9 further corroborates this notion: it includes four additional products (brownie mix, butter, margarine, and chocolate chips) and all of them are baking products.

To this point, we have used CePS simply to explore the neighborhood of individual items, similar to the way in which we might apply association rules networks. As we mentioned before, however, the CePS algorithm is actually much more general, and can handle any number of query nodes. The following discussion explores the ability of CePS to explain a single association rule.

Figure 10 shows a ten-node center-piece subgraph for one of the less intuitive (and more interesting) rules in the dataset: DIET_COKE_20_OZ, YORK_MINT_PATTIES → NEWSPAPER_CHICAGO_TR. Specifically, it is a center-piece subgraph with those three items as query nodes and a budget of 10. The three items in question seem to be entirely unrelated, and yet the rule is ranked highly by a number of interestingness measures (Table 1). Ideally, the center-piece subgraph would illuminate the relationship between the products and explain the association.

Looking at the network, we see something interesting. In addition to patties and Kit Kat, which appear in the association rules network of Fig. 7, we also see three more types of candy: Hershey's, Mounds and Chuckles. This observation implies that there is some sort of relationship between Chicago Tribune, Diet Coke, and candy. As it turns out, the newspapers in our store are located at the front of the store, next to the rack where those candies are sold.

Figure 10 shows that, because center-piece subgraphs can consider the entire product network without requiring excessive computation time or providing overwhelming output, they are very effective for *exploration* or *validation* of relationships provided by other methods. As such, they complement nicely the other techniques outlined in this paper.

Center-piece subgraphs require substantially more parameters than any of the other techniques we have discussed. All of our experiments were conducted on small networks ($b \approx 10$), with "hard AND", meaning that k is equal to the number of query nodes. Though we did not conduct any detailed studies of the parameter selection process, informally we found that the choice of k and b makes little difference in the quality of the subgraph discovered. By choosing b to be large, we observed that popular products such as `BULK_CANDY` and `BAGEL` came to be included in the subgraph. Altering k had no discernible effect for the types of queries we tried.

5.3 A strategy for market basket analysis

The research we present here has allowed us to make and corroborate a number of significant observations about market basket analysis of real-world data. We re-state the chief observations here, citing the work of others where appropriate.

1. Deriving interesting, actionable knowledge from association rules is difficult because rulesets are often muddied by a preponderance of obvious or redundant rules (Klemettinen et al. 1994).
2. One can choose to mine *maximal* or *closed* itemsets instead, but these techniques fail to prune away many redundant rules.
3. Similarly, one may choose to rank rules by an *interestingness* measure, but there are many such measures to choose from and they may rank rules inconsistently (Tan et al. 2004). As such, it may be difficult to choose an appropriate measure in the absence of prior knowledge.
4. Detecting *communities* of products within the network formed by customer purchases can alleviate redundancy by discovering larger, more expressive relationships among groups of products. However, community detection is less effective within the dense *core* of the network and requires a minimum support threshold, which imparts parameter sensitivity.
5. *Association rules networks* are more effective at exploring the core of the network, provided that the chosen target product appears in a large number of association rules. Under other circumstances, they are highly sensitive to the choice of target product and certain networks, even for very popular products, are small and uninformative.
6. *Center-Piece Subgraphs* are useful for explaining or validating relationships discovered by other methods because they do not require a support or confidence threshold to be effective. They are less useful for general analysis because they are necessarily limited in scope.

This list of observations naturally suggests a unified strategy for the analysis of unseen market basket data. First, select a minimum support threshold. On the basis of this threshold, construct a product network and discover communities. The structure of the *interesting* communities in the network [as defined by (5)] provides a quick overview of any especially strong relationships within the data. The discovered relationships are generally more complex and expressive than those discovered with association rules.

Next, the analyst should decide on a minimum confidence threshold and discover association rules. Choosing a popular product, such as the product that appears in the most rules, as the target product, construct an association rules network. This network will provide a roadmap of some of the important relationships within the core of the network and may illuminate some associations that were not clear in the list of communities.

The set of communities and the association rules network, along with the actual list of association rules if desired, will provide a degree of insight into customer behavior in the store. As a final step, one can apply Center-Piece Subgraphs to analyze carefully selected subsections of the entire (unpruned) network. These subgraphs can serve to corroborate or debunk hypotheses about customer behavior or explain unexpected results in the data. Our experiments have suggested that Center-Piece Subgraphs are most effective if the edges of the network are weighted by confidence rather than support.

5.4 Choosing the minimum support parameter

Since the first step in our proposed procedure requires the user to choose a minimum support parameter, we attempt to provide some guidance into this choice. We are aware of no prior work from which to draw, but one can imagine several reasonable options. For example, one might select an arbitrarily high threshold and iteratively reduce it until the number of rules becomes unmanageable. Alternatively, one may attempt to find a certain number (some hundreds or thousands) of rules, or a certain number of rules that score highly based on his or her favorite interestingness measure.

All of these are valid choices and to evaluate them critically is beyond the scope of this work. However, if community detection is the target then existing community detection research affords us another option. In Sect. 3, we briefly alluded to the fact that community detection algorithms find poor communities at low levels of minimum support. This fact can be used, in principle, to choose a minimum support threshold.

Modularity (2) provides us with a measure of the quality of a community structure. It follows, then, that discovering communities at a given support threshold with a modularity-maximization algorithm (e.g. Blondel et al. 2008; Newman 2004, 2006) will provide an estimate of the quality of the communities available at that threshold. This suggests the following procedure:

1. Beginning with a very low support threshold (possibly one transaction), discover communities using a modularity-maximization algorithm.
2. Iteratively increase the threshold until the modularity of the discovered community structure begins to plateau or decrease.
3. If there are several thresholds with very similar modularities pick the lowest one, as it preserves information about the greatest number of products.

Figure 11 shows the modularity of the communities discovered by our implementation of Newman's eigenvector modularity algorithm (Newman 2006) as a function of the minimum support threshold. We chose this algorithm in particular because it is one of the more effective at finding high-modularity decompositions. The graph shows a local maximum at a minimum support of 50 transactions (0.008%) and a global maximum at 110 (0.017%). This suggests that a minimum support threshold of 50 transactions may have been superior to our fairly arbitrary choice of 0.01%. Further evaluation of this method of support tuning will make interesting future work.

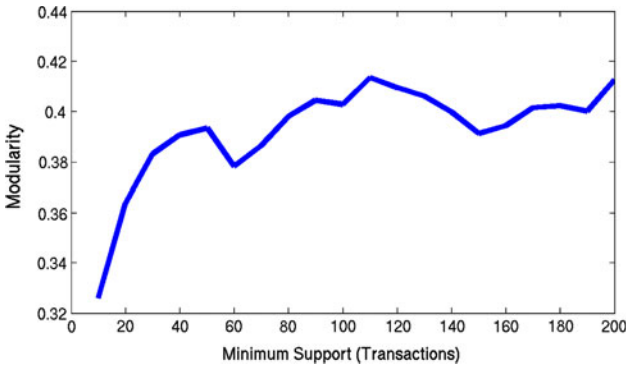


Fig. 11 Modularity of discovered communities as a function of minimum support

6 Related work

Before concluding, we wish to briefly acknowledge a small amount of related work that did not fit cleanly into other parts of the paper. Several authors (e.g. Klemettinen et al. 1994; Hao et al. 2001) use graphs to visualize co-purchases between products. We employ similar techniques to present our results, but claim no originality in doing so.

Clauset et al. (2004) 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. (2001) 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.

Cavique (2007) transforms a transaction database into a graph for the purpose of discovering frequent itemsets. Specifically, the paper employs a heuristic to find maximum-weighted cliques of size *k*, which are then returned as approximate *k*-itemsets. A similar maximum-weighted-clique approach could be applied to discover communities in our product network (see Du et al. 2007), but its asymptotic complexity of $\Theta(n^3)$ is greater than that of the algorithms we have applied. Fonseca et al. (2005) use a graph-based representation of association rules (similar, but not identical, to association rules networks) in order to disambiguate and expand user queries to search engines. For a query term Q, the authors build a directed network of terms where the edge $Q_i \rightarrow Q_j$ exists if the association rule $Q_j \rightarrow Q_i$ holds in the search engine session logs. The strongly connected components in this graph are used to define *concepts* that may be helpful in disambiguating the user's query.

7 Conclusion

This work deals primarily with the application of network techniques to the problem of *market basket analysis*: the location of meaningful associations in customer purchase data. There is an overwhelming abundance of prior research in the mining of mining market basket data in general, and the use of association rules in particular. The bulk of this research has focused on developing algorithms for mining association rules (Agrawal and Srikant 1994; Brin et al. 1997a, b; Zaki et al. 1997; Zaki 1999), techniques for visualizing association rules (Blanchard et al. 2003; Hao et al. 2001; Klemettinen et al. 1994; Wong et al. 1999), techniques for eliminating redundant rules (Klemettinen et al. 1994; Gouda and Zaki 2001; Zaki 2000; Zaki and Hsiao 2002), objective measures of association interestingness (DuMouchel and Pregibon 2001; McGarry 2005; Tan et al. 2004), or comparing the performance of association rule algorithms on either real or synthetic datasets (Hipp et al. 2000; Zheng et al. 2001). However, there has not been much work from a practitioner's view point towards answering: *Given an unseen market basket dataset, what set of steps should I follow to conduct a thorough, complete analysis?* Our work provides a comprehensive framework aimed at answering this question.

First, we study the properties of *networks of products* and show that detecting *communities* within these networks can uncover expressive relationships between products that may be difficult to find with association rules. We show that, in addition to being more expressive than association rules (in that relationships can be expressed more compactly) the structural information available in communities can assist with financial decisions such as the location of profitable promotions. Finally, we develop a novel measure of interestingness for communities of products and show that it favors communities which intuitively seem interesting.

Further, we study the application of two existing techniques, association rules networks (Chawla et al. 2003, 2004; Pandey et al. 2009) and Center-piece subgraphs (Tong and Faloutsos 2006) to the market basket problem. We find that these algorithms complement community detection in the sense that they can be used effectively to find relationships that communities are unlikely to discover. On the basis of this observation, we propose a very general framework for the mining of unseen market basket data in the absence of background knowledge. The framework employs community detection as an initial exploratory step, using association rules networks to uncover relationships within the dense *core* of the network and Center-Piece Subgraphs to validate hypotheses or explore individual relationships that require more explanation.

Acknowledgments This work partially supported by the National Science Foundation under grant NSF 0826958, the NET Institute, and the Arthur J. Schmitt Foundation.

References

Adomavicius G, Tuzhilin A (1999) User profiling in personalization applications through rule discovery and validation. In: Proceedings of KDD. ACM, New York, pp 377–381

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in very large databases. In: Proceedings of the 20th International Conference on VLDB. Santiago, Chile, pp 487–499

Asur S, Ucar D, Parthasarathy S (2007) An ensemble framework for clustering protein-protein interaction networks. In: ISMB/ECCB, pp 29–40

Barabasi A, Bonabeau E (2003) Scale-free networks. Sci Am 288(5):50–59

Blanchard J, Guillet F, Briand H (2003) Exploratory visualization for association rule rummaging. In: KDD-03 workshop on multimedia data mining (MDM-03)

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks

Brijs T, Vanhoof K, Wets G (2003) Defining interestingness for association rules. Int J Inf Theor Appl 10(4):370–376

Brin S, Motwani R, Page L, Winograd T (1998) What can you do with a Web in your Pocket? Data Eng Bull 21(2):37–47

Brin S, Motwani R, Silverstein C (1997) Beyond market baskets: generalizing association rules to correlations. In: Proceedings of the ACM SIGMOD, pp 265–276

Brin S, Motwani R, Ullman J, Tsur S (1997) Dynamic itemset counting and implication rules for market basket data. ACM SIGMOD Record 26(2):255–264

Cavique L (2007) A scalable algorithm for the market basket analysis. J Retail Consumer Serv 14(6):400–407

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

Chawla S, Davis J, Pandey G (2004) On local pruning of association rules using directed hypergraphs. In: 20th international conference on data engineering

Cho Y, Kim J, Kim S (2002) A personalized recommender system based on web usage mining and decision tree induction. Expert Syst Appl 23(3):329–342

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

Clauset A, Shalizi C, Newman M (2007) Power-law distributions in empirical data. axriv, 706

Du N, Wu B, Pei X, Wang B, Xu L (2007) Community detection in large-scale social networks. In: Proceedings of WebKDD. ACM, pp 16–25

DuMouchel W, Pregibon D (2001) Empirical bayes screening for multi-item associations. In: Proceedings of KDD, pp 67–76

Fonseca B, Golgher P, Pôssas B, Ribeiro-Neto B, Ziviani N (2005) Concept-based interactive query expansion. In: Proceedings of CIKM. ACM, p 703

Gouda K, Zaki M (2001) Efficiently mining maximal frequent itemsets. In: Proceedings of ICDM. IEEE Computer Society, pp 163–170

Han J, Pei J (2000) Mining frequent patterns by pattern-growth: methodology and implications. ACM SIGKDD Explor Newslett 2(2):14–20

Hao M, Dayal U, Hsu M, Sprenger T, Gross M (2001) Visualization of directed associations in e-commerce transaction data. In: Proceedings of VisSym, vol 1, pp 185–192

Hipp J, Güntzer U, Nakhaeizadeh G (2000) Algorithms for association rule mininga general survey and comparison. ACM SIGKDD Explor Newslett 2(1):58–64

Kleinberg J, Lawrence S (2001) The structure of the web. Science 294:1849–1850

Klemettinen M, Mannila H, Ronkainen P, Toivonen H, Verkamo A (1994) Finding interesting rules from large sets of discovered association rules. In: Proceedings of CIKM, pp 401–407

Massen C, Doye J (2005) Identifying communities within energy landscapes. Phys Rev E 71(4):46101

Mauri C (2003) Card loyalty. A new emerging issue in grocery retailing. Journal of Retailing and Consumer Serv 10(1):13–25

McGarry K (2005) A survey of interestingness measures for knowledge discovery. Knowl Eng Rev 20(01):39–61

Newman M (2004) Detecting community structure in networks. Eur Phys J B Condens Matter Complex Syst 38(2):321–330

Newman M (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74(3):36104

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

Palmer C, Faloutsos C (2003) Electricity based external similarity of categorical attributes. Lecture notes in computer science, pp 486–500

Pandey G, Chawla S, Poon S, Arunasalam B, Davis J (2009) Association rules network: definition and applications. Statistical analysis and data mining 1(4)

Steinhaeuser K, Chawla N (2008) Community detection in a large-scale real world social network. In: LNCS. Springer, Berlin

Tan P, Kumar V, Srivastava J (2004) Selecting the right objective measure for association analysis. Inf Syst 29(4):293–313

Tong H, Faloutsos C (2006) Center-piece subgraphs: problem definition and fast solutions. In: Proceedings of KDD. ACM New York, pp 404–413

Tong H, Faloutsos C, Pan J (2006) Fast random walk with restart and its applications. In: Proceedings of ICDM, pp 613–622

Wong P, Whitney P, Thomas J (1999) Visualizing association rules for text mining. In: 1999 IEEE Symposium on Information Visualization, 1999 (Info Vis' 99) Proceedings, pp 120–123

Xiong H, Tan P, Kumar V (2006) Hyperclique pattern discovery. Data Mining Knowl Discov 13(2):219–242

Zaki M (2000) Generating non-redundant association rules. In: Proceedings of KDD. ACM New York, pp 34–43

Zaki M, Hsiao C (2002) CHARM: An efficient algorithm for closed itemset mining. In: 2nd SIAM International Conference on Data Mining, pp 457–473

Zaki M, Parthasarathy S, Ogihara M, Li W et al (1997) New algorithms for fast discovery of association rules. In: Proceedings of KDD, vol 20

Zaki MJ (1999) Parallel and distributed association mining: a survey. IEEE Concurr 7(4):14–25

Zheng Z, Kohavi R, Mason L (2001) Real world performance of association rule algorithms. In: Proceedings of KDD. ACM, New York, pp 401–406