Cut-Cliques in a Market Network

Pedro Martins, Antonio Ladrón de Guevara, Helena R. Lourenço
Faculdade de Ciências da Universidade de Lisboa
Instituto Superior de Contabilidade e Administração de Coimbra
Department of Economics and Business, UniversitatPompeuFabra
Email (contact author): antonio.ladron@upf.edu

Abstract

This paper focuses on the application of cliques into market networks, addressing marketing related problems, where the nodes of the network represent a set of products and the links represent pairs of products appearing in the same market basket. In this network, a clique characterizes a subset of products found together in common market baskets. Instead of searching for the largest set of products in a clique, as is most usual in the literature, we search for a clique (of any size) with the largest number of links to other products in the network. This is a recently proposed clique's related problem, being designated as maximum edge neighborhood clique or maximum cut-clique. This new structure is expected to reveal customer's choices when putting products to a market basket, and it can be seen as another contribution for mining a market network. We propose iterated local search based heuristics for this newly problem and explore its application in real world market networks.

Keywords:OR in marketing, clique's edge neighborhood, cut-cliques, iterated local search algorithms, market basket, data mining.

Introduction

Searching for dense components in a network has long been attracting many researchers from different areas. Among those structures there is the concept of a clique, in which all elements are pairwise adjacent. This structure is expected to reveal a strongly related set of elements.

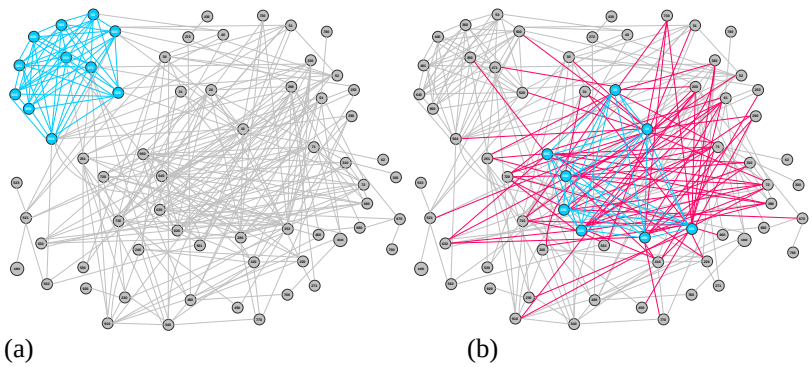
A large number of applications involving cliques were discussed in the literature since long. Some of those applications can be found in coding theory, fault diagnosis, computer vision, pattern recognition (see, e.g., Bomze et al. 1999), telecommunications, marketing, financial markets, social networks and in molecular and biological networks (see, e.g., Butenko and Wilhelm 2006; Strickland et al. 2005; and Cavique 2007), among others.

To formalize the problem, let $G=(V,E)$ be an undirected graph, where $V=\{1,\ldots,n\}$ is the set of nodes and $E\subseteq V\times V$ the set of edges. A clique of G is a subset of nodes $C\subseteq V$ whose elements are pairwise adjacent, that is, $(i,j)\in E$ for all pairs $i,j\in C$. Finding the maximum cardinality clique in G is known as the Maximum Clique (**MC**) problem.

In the present paper we consider a different clique's related problem. Instead of searching for the largest size clique in the graph, we want a clique (of any size) with the largest number of edges incident to the nodes in the clique, besides those within the clique. This problem has been recently introduced in Martins (2012), where formulations were proposed and showed its applicability to some real world problems. The clique with the largest number of edges in the neighborhood of the clique is known as *maximum cut-clique* (**MCC**) *problem.* Figure 1-(a) shows a maximum clique (in blue) and Figure 1-(b) shows a maximum cut-clique (in red) in a 68 nodes graph. The maximum clique solution includes 11

nodes (in blue), while the maximum cut-clique solution uses an 8 nodes clique (in blue) with 93 edges in the cutset$E'(C)$ (edges in red). Note that, the total number of edges in the neighborhood of the clique in (a) is equal to 22, which is much smaller than the edge neighborhood of the clique in (b). Yet, the smaller sized clique (in b) is much more engaged in the network than the largest size clique (in a), which may suggest that the smaller sized clique can be more relevant within the whole network. In fact, searching for a clique of largest size can lead us into an isolated component of the graph, being displaced from the "crowdie" zone.

Figure 1: (a) a maximum clique (nodes/edges in blue), and (b) a maximum cut-clique (edges in red).



The motivation to focus on the Maximum Cut-Clique Problem arrives from the application to Market Basket Analysis. The main objective of this field is to analyze large dataset of store transactions and obtain relevant insights to do a better planning of the marketing actions and strategies. This work presents new and innovative techniques for mining information on market networks. We believe that the information returned by a maximum cut-clique in a market network can bring valuable insights from customers' habits.

In the next section we review briefly the literature in Market Basket Analysis, relate this one with the approach proposed in this work and detach the practical relevance of identifying clique's edge neighborhood in a market network. The application of these methodologies to a real-world market network is discussed in section 3. The paper is closed with a conclusions section.

Motivation for using clique's edge neighborhood in a market network

In this section, we briefly review the literature on basket analysis, relate it with the cut-clique approach proposed here and explain the expected impact of this methodology in the analysis of product networks.

In marketing, the field of Market Basket Analysis consists in identify meaningful associations in a customer transaction dataset. The area is becoming every day more relevant due to the increasing amount of data on household purchase history available formanufacturers and retailers today, from loyalty cards to household panels. The information obtained from the analysis of this data can have an important impact in the business strategy and operations, for example product placement, optimal product-line offering, personalized marketing campaigns and product promotions. The main objective is the search of meaningful associations in customer purchase data, as for example if a customer buys product A is very likely to buy product B. In the literature there are different approaches to find the relevant relationships among the products, most of them based on data mining techniques. Some of the

common techniques are: Association Rules, Detecting Communities, Association Rules Networks, Hyperclique Pattern Discovery and Center Piece sub-graphs. For a survey on Market Basket Analysis see Hipp et al. (2000), Raeder and Chawla (2011) and Zaki (1999).

The traditional and oldest interdependence approach for analyzing market basket data is the detection and estimation of conditioned purchase probabilities for pairs of products purchased in the same basket, known as association rules. However, consumer datasets frequently contain hundreds of association rules, so filtering and selecting the relevant subsets of interdependence patterns is not a trivial or easy task, Klemettinen, et al. (1994). In the last few years researchers have developed several techniques to address this important limitation of the traditional approach (see Raeder and Chawla 2011 for a detailed description). Among these methodologies, the analysis of cliques or network-based rules allows for finding relevant and meaningful relationships across sets of products in large consumer purchase datasets.

In this research wedescribe the application of the cut-clique model and corresponding algorithm to a category of products to obtain relevant information on the household purchase behavior.Next we describe the basket network and the implications of applying the maximum cut-clique problem in market basket analysis. For the empirical application we consider the purchase history of individual households during a given period of time for a certain product category presenting multiple purchases. Today this data can be easily obtained, for example from loyalty cards orhousehold panels. Hence, each node of the network represents a product and an edge (i,j) in the network indicates that there was a customer that bought together the two products i and j in one of his/her visits to the store(s). For a set of products strongly linked (those in the clique) its edge neighborhood indicates the chance of buying a product outside the clique by someone that bought a product in the clique. Looking for the clique with largest edge neighborhood should reveal the set of products (clique) with strongest chance of attracting other products in their vicinity. Finding the clique with largest incidence to its neighborhood is the motivation for the present paper for mining a given market network.

We argue that by modeling the market basket data as a product network, and obtaining the cut-cliques of this network, interesting association rules, expressive clusters and clusters of products with great influence can be discovered and the object of further marketing analysis. The proposed approach is able to generate insights and analysis difficult to find with other more traditional methods. Finding a maximum cut-clique on a given graph contributes for mining information on a given market network. The idea is to isolate communities of products within the network with the larger number of relationships with other products in the store. A community is a group of products that are strongly connected between them. The detection of the communities is a particular relevant area in Market Basket Analysis. In this case we look for a clique, a complete sub-network, with the larger number of links with the rest of the network. The number of links can be seen as a measure of the utility of the communities, and the identification of this cut-clique can provide information that can be useful to take marketing actions, as for example product line decisions, joint promotions, and product placement to increase the sales.

The possible impact of applying the maximum cut-cliques to a market network can be very different, depending on the definition of the basket network. In the general case, where the nodes are all products and the edges represent at least one transaction with both products in the basket in a period of time, the implication could be identifying the basket with more relationships outside the clique and consequently, the highest cross-product potential. A similar analysis can be obtained if it is used a network with a threshold, which in this case the implications are stronger since they represent at least a minimum number of transaction with similar characteristics. If we identify nodes representing products of the same brand, and the edges as before, the analysis could be useful to identify variety-seeking behavioral patterns

among brand-loyal consumers. Finally, if the analysis is run at the individual household level, the maximum cut-clique analysis could provide insight on personalized marketing, for example send a promotion e-mail or SMS on the products of the neighborhood of the clique to increase the sales.

Raeder and Chawla (2011) say "... no techniques currently available in the literature sufficiently addresses the problem of finding meaningful relationships in a large transaction databases.". With the present work we do not expect to solve this important issue in marketing, but we believe we make a strong contribution to obtain relevant and important information from a market basket database. We proposed an innovative technique based on cut-cliques, contributing to a new and innovative approach to obtain insights and relevant associations rules from a market network. The techniques proposed here can be complementary to the ones based on data mining, since both provide relevant insights about the market basket. One of the main advantages is that the methods can solve very large scale datasets, and so we can apply the methods to individual SKUs, not only to brands of products.

Maximum cut-cliques applied to a market network

We have applied the proposed methods to a database of transaction data collected with home scanners.[1] We use a household panel database for the British ice cream market containing information over a 2½-year period (January 2006 to June 2008) among 142 households. The dataset contains information for a total of 4,899 items purchased, averaging 1.64 ice cream products per basket, chosen from a total of 691 different *varieties of products* (SKUs)[2] available in the British market.Considering the database, let's define a basket as the set of products purchased in the same day and the same shop by a household. We form a basket for each group of products that have the same household, shop, and purchase date.The analysis has been conducted at the household level, independent of the family member who made the purchase.

The number of products (SKUs) in the network is 691, and the number of edges is 1181. Just to remember, if two products (i and j) are found in the same basket, then the network includes edge (i,j), otherwise, if the two products are found in no common basket, then edge (i,j) doesn't belong to the network.

We have determined the maximum clique (MC) and the maximum cut-clique (MCC). For the solution of these problems we have used iterated local search (ILS) based heuristics.These algorithms are derived from the Dynamic Local Search methods described in Pullman and Hoos (2006). The detailed description of these methodologies is beyond the objective of this research. However, the mathematical formulations, the iterated local search algorithms, and the computational tests addressing the problems are available from the authors upon request.

We now describe the results, the implications for the analysis of the product interactions, and the managerial implications. The largest clique in the network (maximum clique) includes the following 8 products:

prod id number	prod description	# external links
147	FRDRKS DARK CHOC ICE VNLA 10PK	4
148	FRDRKS CHOC ICE NPLTN 10PK	13
149	FRDRKS LGHT CHOC ICE VNLA 10PK	3
375	CDBRY CONE DRY MLK MINT 4PK	9
489	DLMNT LLY RSPBRY SMOOTHIE 3PK	16
518	NSTL LOLLY ROLO STCK 6PK	0
539	FRDRKS CHOCOLATE 2LT	0
541	FRDRKS STWBRY 2LT	0

[1]See Brennenberg et al (2008) for a detailed description of IRI home scanner data base.
[2]Stock Keeping Unit.

This solution indicates that there are baskets with all possible pairings among these 8 products. This is the largest set of products having this property. Together, these products have 45 links to the remaining products in the network, which means that, besides the clique's baskets, each one of the 8 products share the same basket with other products in 45 occasions. Those other adjacent products involve 32 items.

The maximum-clique analysis allows for the identification of the biggest set of items commonly purchased in conjunction with others at least once in a same basket. The analysis of the product network reveals some interesting results. As explained previously, the cliques constitute groups of products that have been bought together in a basket with all the other products of the set. As for many product categories, a significant segment of households buy more than one ice cream in the same basket. Identification of cliques allows for determining the attributes or dimensions in which multiple-purchase households seek variety. For instance, five out of the eight products constituting the maximum clique are big formats - 2 liters or 10-unit packages - of an Italian-style luxury brand, Fredericks's, varying in the following five flavors: Chocolate, Strawberry, Vanilla ice cream with dark chocolate, Vanilla ice cream with light chocolate, and Neapolitan ice cream with chocolate. The remaining three products of the basket are also multiple-unit packages of three products from leading manufacturer brands: Del Monte, Nestlé, and Cadbury.

Noteworthy, the set of products constituting the maximum clique is not the clique with largest incidence to other products in the network. In fact, the set of products forming a clique and with the largest number of links to the remaining products in the network (maximumcut-clique) involves only 6 products, but having 100 links to the remaining products. Those 6 products are:

prod id number	prod description		# external links
21	WALLS BL RBN VNLLA	2LT	10
65	WALLS MINI MILK LOLLIES	12PK	14
66	MARS CHOC ICE	4PK	23
72	WALLS MAGNUM WHITE STCK	3PK	23
80	NSTL LOLLY FAB	8PK	21
305	WALLS MNI TWISTER STW+LMN	8PK	9

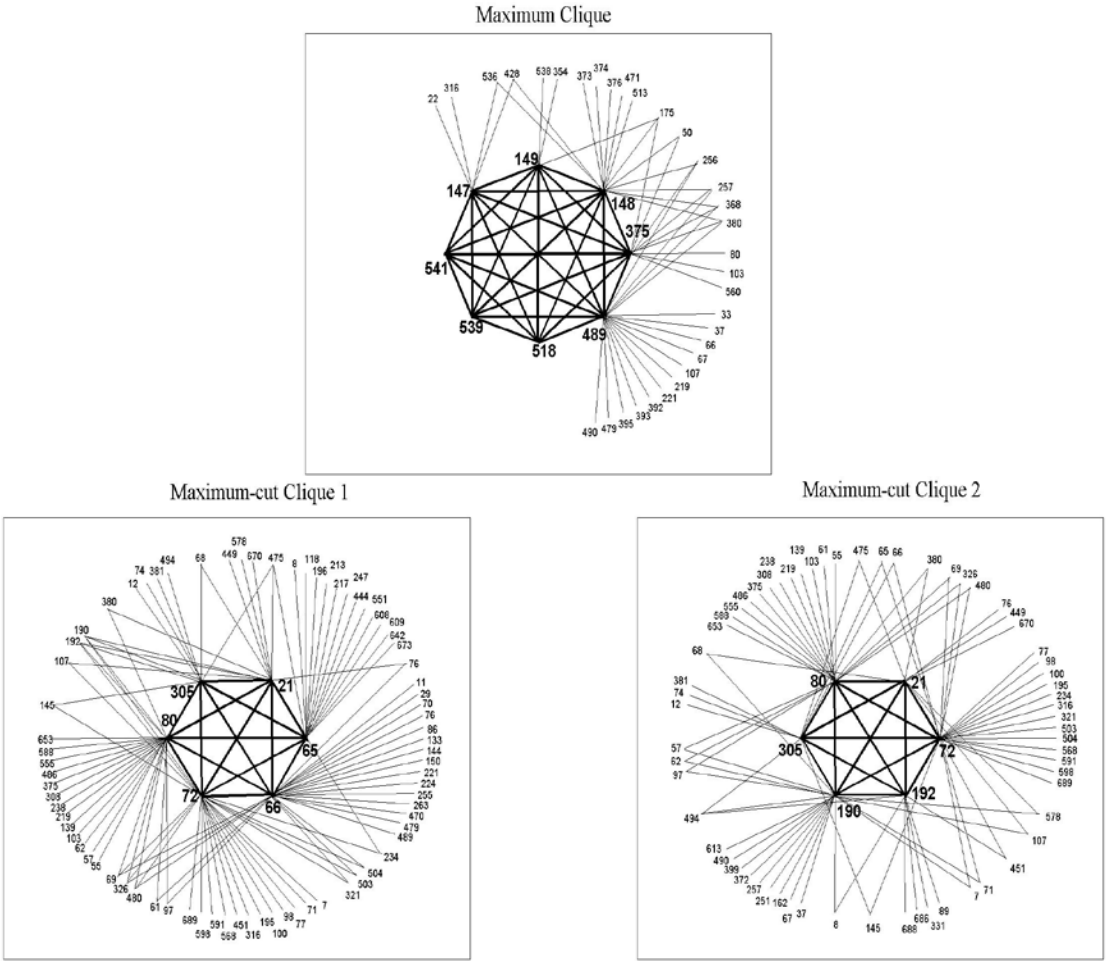
These 6 products form a clique, which means that all pairs of products in this set are found in householder's baskets. They are adjacent to 75 other products in the network, involving 100 links, which stresses its strong engagement. So, probably, most of the householders buying these 6 products are also strong potential buyers for the remaining products, especially those products involved in the 100 links.

The second-best solution, with a smaller number of links (just 93) is:

prod id number	prod description		# external links
21	WALLS BL RBN VNLLA	2LT	10
72	WALLS MAGNUM WHITE STCK	3PK	23
80	NSTL LOLLY FAB	8PK	21
190	WALLS CORNETTO STWBRY	6PK	20
192	WALLS CORNETTO CHOC N NUT	6PK	10
305	WALLS MNI TWISTER STW+LMN	8PK	9

In both solutions, the maximumcut-clique analysis provides smaller sets compared to the maximum clique solution. However, we can also see that in both solutions the set of products obtained when maximizing the external interactions is formed by the top-selling ice creams from leading brands: Cornettos (Walls), Magnums (Walls), Blue Ribbon (Walls), Lollyfabs (Nestlé) and Mars chocolate bar (Mars). Figure 2compares the neighborhoods of the previously discussed cliques.

Figure 2: Comparing the neighborhoods of the three cliques.



Noteworthy, while the maximum-clique analysis focuses on the biggest set of interacting products independent of the number of purchases, some items in the set may have been bought just a few times, or even only once. Therefore, the maximum-clique analysis may not be the most appropriate technique to identify and analyze representative product interactions in the dataset. However, the maximumcut-clique analysis identifies the set of products with the maximum number of links with the other products of the network, revealing interacting patterns from leading-sale products. A visual comparison of the biggest and second biggest maximumcut-cliques with the maximum clique reveals significant differences in the set of products forming the sets. While the average number of purchases of the 8 products forming the maximum clique is 9.3, the average number of purchases of the 6 products forming the biggest and second biggest maximumcut-cliques are 34.2 and 38.5 respectively (see Table 1 for a comparison of the different characteristics of the networks.

Table 1: Descriptive values for the maximum clique and maximum cut-clique solutions.

Solution	Number of products	Number of external products	Number of external links	Average number of purchases per product
Maximum clique	8	32	45	9.3
Maximumcut-clique 1	6	75	100	34.2
Maximumcut-clique 2	6	63	93	38.5

Conclusions

In this work we consider the Maximum Cut-Clique problem, proposing an Iterated Local Search heuristics to solve it (R-ILS and D-ILS algorithms). We also describe the importance of the application of cliques into market networks, addressing marketing related problems in the field of market Basket Analysis. The information obtained from the analysis based on the Maximum Cut-Clique can have an important impact in the business strategy and operations, for example product placement, optimal product-line offering, personalized marketing campaigns and product promotions.

We present the application of the algorithms to a household panel database for the British ice cream market, and analyze the results and conclude the importance of applying these techniques to obtain relevant information from consumer's databases.

This work open a new line of research related with the application of Clique based models to evaluate market baskets and obtain different type of information from the traditional approaches in marketing. Future work is oriented in applying these models and techniques to larger databases on several different products and markets. An important extension is to consider a network with weights in the edges, and look for the maximum weight cut-clique.

References

Bomze, I.M., Budinich, M., Pardalos, P.M., Pelillo, M., 1999.The maximum clique problem. In: Du, D.Z., Pardalos, P.M. (Eds.), Handbook of combinatorial optimization (suppl. Vol. A). Kluwer, Dordrecht, pp. 1-74.

Butenko, S., Wilhelm, W.E., 2006.Clique-detection models in computational biochemistry and genomics.European Journal of Operational Research 173, 1-17.

Cavique, L., 2007. A scalable algorithm for the market basket analysis.Journal of Retailing and Consumer Services 14, 400-407.

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.

Martins, P., 2012. Cliques with maximum/minimum edge neighborhood and neighborhood density. Computers & Operations Research 39, 594-608.

Pullan, W., Hoos, H.H., 2006.Dynamic local search for the maximum clique problem.Journal of Artificial Intelligence Research 25, 159-185.

Raeder, T., Chawla, N.V., 2011.Market basket analysis with networks. Social Network Analysis and Mining 1 (2), 97-113.

Strickland, D.M., Barnes, E., Sokol, J.S., 2005. Optimal protein structure alignment using maximum cliques.Operations Research 53 (3), 389-402.