Overlap Versus Partition: Marketing Classification and Customer Profiling in Complex Networks of Products

Diego Pennacchioli[1,2], Michele Coscia[3], Dino Pedreschi[4]

[1] IMT - Lucca, P.za San Ponziano, 6, Lucca, Italy, diego.pennacchioli@imtlucca.it
[2] KDDLab ISTI-CNR, Via G. Moruzzi, 1, Pisa, Italy, diego.pennacchioli@isti.cnr.it
[3] CID - Harvard University, 79 JFK Street, Cambridge, MA, US, michele_coscia@hks.harvard.edu
[4] KDDLab University of Pisa, Largo B. Pontecorvo, 3, Pisa, Italy, pedre@di.unipi.it

***Abstract*—In recent years we witnessed the explosion in the availability of data regarding human and customer behavior in the market. This data richness era has fostered the development of useful applications in understanding how markets and the minds of the customers work. In this paper we focus on the analysis of complex networks based on customer behavior. Complex network analysis has provided a new and wide toolbox for the classic data mining task of clustering. With community discovery, i.e. the detection of functional modules in complex networks, we are now able to group together customers and products using a variety of different criteria. The aim of this paper is to explore this new analytic degree of freedom. We are interested in providing a case study uncovering the meaning of different community discovery algorithms on a network of products connected together because co-purchased by the same customers. We focus our interest in the different interpretation of a partition approach, where each product belongs to a single community, against an overlapping approach, where each product can belong to multiple communities. We found that the former is useful to improve the marketing classification of products, while the latter is able to create a collection of different customer profiles.**

I. Introduction

Due to the advancements in data storage, computational power and data analysis techniques, in the last years authors in many disciplines have shifted their attention from theoretical models to data-driven problems. This is the so-called "Big Data" paradigm, in which the vast amount of data, previously unavailable, has uncovered novel problem definitions and enlarged the set of hypotheses and theories that can be tested.

In this paper we focus in particular on the augmented quality and quantity of information that we can collect about customer behavior. Using the tracking abilities of customer fidelity cards, a supermarket chain can analyze the behavioral patterns of its customers: where do they come from? What combination of products are frequently bought by certain customers? Is it possible to quantify the diversity of needs of different customers? These and many other questions can now be answered.

As a consequence, a number of applications and techniques have been applied to customer segmentation and analysis. For example, some authors developed a framework to predict in which shop a customer will buy a given product, given the customer's residence and its degree of sophistication [1].

A particular fruitful technique applied to big data and customer behavior is complex network analysis. Two products can be connected if they are frequently co-purchased by the same customers, allowing a complex structure of products to emerge [2]. The topological properties of this complex structure are informative about how customers perceive product relations, just like the collaborative filtering of Amazon and Netflix, but on a broader product typology set. For instance, sets of products may be very densely connected the one with the other, because customers always buy them together.

The task of finding sets of nodes densely connected in a complex network is known as "community discovery" [3]. Community discovery is one among the most prolific subbranches of complex network analysis. Hundreds of papers have been written on the subject, and dozens of algorithms have been proposed to solve it. As a result, the scientific community has come to agree that there is no unique solution to community discovery, given the many different possible definitions of "community" that can be accepted in different applications.

In particular, one of the most important distinction between community discovery algorithms is about a node's membership to a community [4]. In some methods, a node is forced to belong only to the community it is closest to. This is a partition approach to community discovery (often called "hard clustering", or "disjoint community discovery"). In other methods, a node may be free to join as many communities as necessary. If an algorithm allows this type of output, then it is said to return overlapping communities (also known as "soft clustering" or "community coverage").

Historically, overlapping algorithms were developed as a critique to the partition approach. The theory was that "actual communities" are overlapping, and therefore the partition approach was obsolete. Our point is that the two approaches are not mutually exclusive. In the same context, they yield different results because the problem they address is different

and there is no "better" or "worse" method.

The aim of this paper is to investigate the typology of results that different approaches to community discovery can achieve while analyzing a complex network of products. When applied to a network created connecting products if they are co-purchased by the same customers, community discovery will return groups of products that are "related" to each other. Our aim is to understand what "related" means under different community definitions, in particular when our aim is to find a community partition *vis-à-vis* when our aim is to find an overlapping community coverage.

To prove our point, we collected data about more that 24 thousand products, co-purchased by a million customers in more than 80 millions shopping sessions from four regions in the center of Italy. Using their purchases, we created a product-product network connecting products if they were co-purchased during the same shopping session. We then applied two state-of-the-art community discovery algorithms on this structure: one yielding a disjoint community partition (Infomap [5]), the other yielding an overlapping community coverage (Hierarchical Link Clustering [6]).

Our results confirmed that the different community definitions returned two very different sets of results. We observed that there is no clear ranking in the quality of these results, i.e. there is no clear way to determine which algorithm performed "better". On the other hand, both the partition and the overlapping approach returned results that can be utilized for different tasks. The disjoint communities proved to be useful for the redefinition of the product marketing classification. The overlapping communities, instead, represent specific customer behaviors, and therefore provide useful data for the task of customer profiling.

To sum up, the contributions of the paper can be summarized as follows:

- To the best of our knowledge, this is the first empirical test able to provide an insight about the practical usefulness of different community discovery approaches in a real-world analytic scenario, in particular about the difference between a partition approach versus an overlapping approach. As a consequence,
- We showed how partition-based community discovery is useful as a novel approach to the construction of marketing, and possibly general purpose, classifications;
- We provided a novel approach to customer profiling, via overlapping community discovery of product co-purchase networks.

The rest of the paper is organized as follows. We present relevant literature in Section II. Section III is dedicated to the description of our dataset. The partition and overlapping community discovery methods we used in this paper are discussed in Section IV. Section V contains our case study. Finally, Section VI concludes the paper with future works.

II. Related Works

This paper brings together two different branches of research. On one hand, it is a paper about analyzing customer behavior with computer science techniques, in particular data mining and complex network analysis. On the other hand, it is focused on a specific branch of network analysis, namely community discovery. Thus, in this section we frame the paper in these two branches, starting with the computer science analysis of markets.

Markets are complex systems, where customers, manufacturers, goods, services, etc., are strongly related each other. For this reason, classical data mining approaches (either focused on one dimension, e.g. classifying the customers; or on direct relations, e.g. co-appearance of a set of products in transactions) are often not sufficient. Hence, during the last years, the attention moved to modeling market environments as complex networks. The entities that can be modeled are diverse: in [7], the authors build a network connecting customers based on communication frequency, to plan a better strategy for targeted marketing in telecommunication services. In [8], [1] the authors use a bipartite graph (a graph where nodes belong to two different classes, in that case customers and products), describing the whole retail market and finding a general law driving the customer behavior. One first attempt to build a network of products is in [2], where the authors use a dataset coming from an university store over the time span of a year (660K transactions, 2200 products). Authors build a directed network (where relationships are not symmetrical) connecting product A to product B if B is frequently purchased when A is purchased. Their approach is based on association rules discovery [9]. Here the authors are interested not at the composition of the communities, but at their overall quality, measured with an aggregation of the confidence attached to the edges.

Moving to the other central aspect of the paper, one classical problem definition in complex network analysis is how to detect functional modules in the network. This is usually known as "Community discovery", borrowing from the social network literature [3]. Given the high relevance of this branch of studies for this paper, we examine this problem with higher depth in Section IV. In general, community discovery is a very popular research field in complex network analysis, with hundreds of papers on the topic and an almost equal number of developed algorithms [4]. The amount of relevant literature is due to the lack of a proper and unique definition of "community" [10] and the potential high impact of research in the field [11]. Historical approaches such as defining a particular quality function (like modularity) for community discovery or the detection of semi-cliques in the network (the k-clique percolation algorithm) have been widely used but are of no interest here given their theoretical downsides and/or their inability to scale for large networks [3]. Successful modern approaches can be divided in two classes, that we will explore in depth in Sections IV-A and IV-B: partition-based algorithms such as Infomap [5], agent-based [12] or label propagation based [13]; and overlap-based like DEMON [14], Hierarchical Link Clustering [6] and overlap label propagation [15], [16].

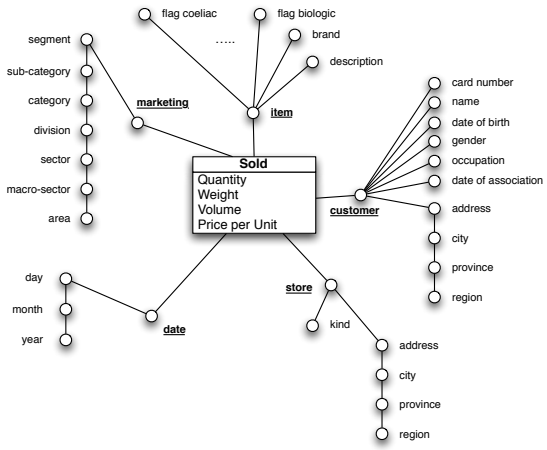


Fig. 1: The Conceptual Data Model (star schema) of the Data Warehouse

III. DATA

The dataset we used is the retail market data of Coop, one of the largest Italian retail distribution company. The conceptual data model of the data warehouse storing the retail data is depicted in Figure 1.

The whole dataset contains retail market data in a time window that goes from January 1st, 2007 to December, 31st 2011. The active and recognizable customers in that interval are $1,066,020$. A customer is active if she has purchased something during the data time window, while she is recognizable if the purchase has been made using a membership card. The 138 stores of the company cover an extensive part of Italy, selling $345,208$ different items.

Each data entry contains information about a product item bought by a customer in a specific store in a specific moment. While most of the dimensions of the data warehouse are clearly understendable, of particular interest is the *Marketing* category. This is used to classify products: it is organized as a tree and it represents a hierarchy built on the product typologies, designed by marketing experts of the company (see Figure 1 for a list of hierarchy levels). The top level of this hierarchy is called "Area" and it is split in three fundamental product areas: *Food*, *No Food* and *Other*. The bottom level of the marketing hierarchy, the one directly on top of the leaves of the tree, is called *Segment* and it contains $7,003$ different values. Each item has a classification in this hierarchy and, thus, we can exploit such tree to choose the most suitable level of aggregation of products. The main difference between *item* level and *Segment* level consists in packaging, size and brand. For example, the three items: half-liter Sugar Free Coca Cola bottle, $6X1.5$ liter Sugar Free Coca Cola bottle's box, and two liters Pepsi Cola bottle, belong all at the same Segment (Sugar Free Cola Drinks).

As claimed above, the dataset contains information about $345,208$ different objects sold in the shops. The main bias introduced by this huge cardinality of products are two:

1) we have information about products that are meaningless for our purposes (e.g. shoppers, discount coupons, etc.), and
2) due to some exception, we have distinct products that have the very same semantic (e.g. 6-bottles regular Coca Cola box and 6-bottles regular Coca Cola box with Santa Claus in the package for Christmas time).

We solved (1) by filtering data using semantic information in the marketing hierarchy. We solved (2) by including in our analysis dataset at most the top 5 sold products (or less, if there are not enough products) for each marketing Segment. Notice that, for each item exception, there always is a product that is top seller over the others with the same semantic. After this filtering phase, the dataset contains $26,862$ different items, that are the nodes of our network, belonging to $5,510$ marketing Segments.

We now want to connect these nodes, to create the product complex networks. Given the big amount of data considered, almost all products have been sold at least once with all the other products. We need to filter out these connections, to focus only on relevant and significant relationships. To this end, we discovered all possible pairs of products sold together (using Apriori [9]), and we calculated, for each of them, the *lift* measure, defined as:

$$\mathrm{lift}(X,Y) = \frac{\mathrm{supp}(X,Y)}{\mathrm{supp}(Y) \times \mathrm{supp}(X)}$$

where X and Y are products, and $\mathrm{supp}(i)$ is the number of baskets containing the item i divided the total number of the baskets in the dataset. $\mathrm{supp}(i)$ is the "Relative Support" of i, i.e. the observed likelihood of having i in a basket. Lift measures how much a pair of items is interesting, calculating how its distribution is related with the distribution of the single items. If lift is equal to 1 we are under the hypothesis of stochastic independence, and the greater lift is, the greater the occurrence rate of the pair is significant.

Since lift is a relative measure, we need also to take under control the popularity of the products composing the pair. In fact, the supports of the single items composing the pairs are in the denominator of the lift formula, and multiplying each other. This implies that the smaller the supports are the greater the lift is inflated, by exaggerating the relevance of products rarely sold and thus not really meaningful. For this reason, we also use the "Absolute Support" of the pair, measuring how many are the occurrences of the couple in the dataset, i.e. the number of baskets containing both products.

To sum up, lift tells us how interesting the pair occurrence is, the absolute support tells us how relevant the pair occurrence is. A pair to be included in our network has to be interesting and relevant at the same time. In Tables I and II we show the cardinalities of the edge set and the node set of different product networks, built using different thresholds on absolute support (in rows) and lift measure (in columns) of the pairs.

To obtain a manageable network, with edges representing associations not very infrequent but strongly reliable, we chose

	1	2	5	8	10
10	20,042,602	11,784,927	1,962,699	825,644	577,954
50	9,141,753	5,356,930	640,933	264,156	187,341
100	5,874,000	3,433,601	376,367	160,049	115,033

TABLE I: Number of edges in the Product Network after filtering with Minimum Absolute Support (rows) and lift (columns).

	1	2	5	8	10
10	16,910	16,769	16,152	15,402	14,949
50	12,268	12,035	11,401	10,797	10,392
100	10,578	10,347	9,734	9,158	8,784

TABLE II: Number of nodes in the Product Network after filtering with Minimum Absolute Support (rows) and lift (columns).

to set the minimum Absolute Support at 10 and the minimum lift at 10. The resulting product network contains $14,949$ nodes and $577,954$ edges, and that is the network we use hereafter, for our case study in Section V.

IV. Community Discovery

It has been described in literature that many real world networks have a non-homogeneous topological distribution of their links [4]. In other words, there are portions of the network with a high edge density and they are usually isolated by other areas with a low edge density. In literature, it has been decided to call "communities" the collections of nodes densely connected one to the other. The task to efficiently detect the communities in complex networks has been called "community discovery".

This branch of network science is very prolific, with hundreds of papers proposing new approaches to the detection of network communities [3]. Given the extensive attention on the subject, two issues have been deeply studied. The first is the notion that there is not a single best method to extract communities from complex networks. It is possible to define "communities" in different ways and different approaches are more or less efficient for a particular community definition [3].

The second issue is the one at the center of investigation of this paper. The assumption that communities are dense subsets of nodes isolated from the rest of the network has been questioned. There is a growing evidence that communities are not really isolated from the rest of the network, but rather overlap the one with the other, sharing nodes. Given this issue, two mutually exclusive approaches to community discovery can be implemented: the partition approach, that follows the main assumption here presented; and the overlap approach, that allows nodes to be classified in more than one community.

It is one of the assumption of this paper that both approaches can yield enlightening, and different, results even in the very same network. For this reason, we briefly present the two community discovery algorithms, that we use in our case study of analysis of a co-purchase product network. In Section IV-A we present Infomap, a community discovery algorithm employing the partition approach; and in Section IV-B we describe the Hierarchical Link Clustering (HLC), an overlap community detector. Both algorithms are non-parametric, i.e. they maximize an internal quality function and return optimal clusters, thus their results are not dependent on our choices. We are aware that many other algorithms with different properties exists in literature, but we consider only these two due to lack of space. In any case, these are two of the best performing algorithms available in literature. We leave a more comprehensive study for future work.

A. Partition Approach

As we discussed, in the partition approach the main assumption is that densely connected nodes are separated from the rest of the network by nodes with sparser connections. We show a simplified example of this assumption in Figure 2(a). In Figure 2(a) we clearly need a partition of the graph, separating nodes 0 to 4 from nodes 5 to 9 and from nodes 10 to 14. As a consequence, algorithms seeking a node partition have to minimize the number of edges between communities, while maximizing the number of edges inside communities. Many non-trivial measures have been proposed. One of the proved most successful is the compression factor allowed by the partition, and it has been proposed in the Infomap algorithm.

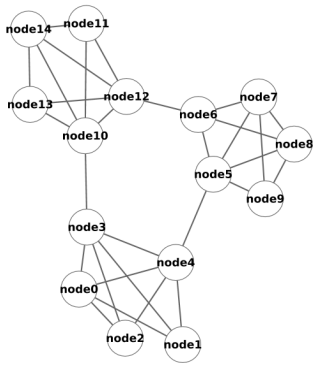
The Infomap algorithm [5] is based on a combination of information theoretic techniques and random walks. It uses the probability flow of random walks on a graph as a proxy for information flows in the real system and decomposes the network into clusters by compressing a description of the probability flow. The algorithm looks for a cluster partition M into m clusters so as to minimize the expected description length of a random walk.

In Figure 2(b) we have depicted the same example of Figure 2(a) where the edge width is proportional to the amount of redundant information shared by the two connected nodes.

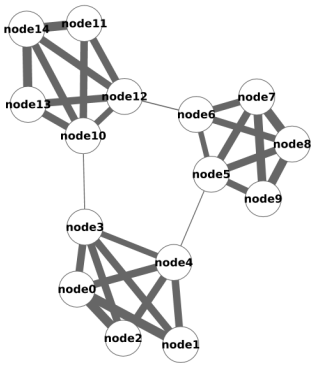
The intuition behind the Infomap approach for the random walk compression is the following. The best way to compress the paths is to describe them with a prefix and a suffix. Each node that is part of the same cluster M of the previous node is described only with its suffix, otherwise with prefix and suffix. Then, the suffixes are reused in all prefixes, just like the street names are reused in different cities. The optimal division in different prefixes represent the optimal community partition. We can now formally present the theory behind Infomap. The expected description length, given a partition M, is given by:

$$L(M) = qH(Q) + \sum_{i=1}^{m} p_i H(P_i).$$

$L(M)$ is made up of two terms: the first is the entropy of the movements between clusters and the second is entropy of movements within clusters. The entropy associated to the description of the n states of a random variable X that occur with probabilities p_i is $H(X) = -\sum_1^n p_i \log_2 p_i$. In (1) entropy is weighted by the probabilities with which they occur in the particular partitioning. More precisely, q is the probability that the random walk jumps from a cluster to

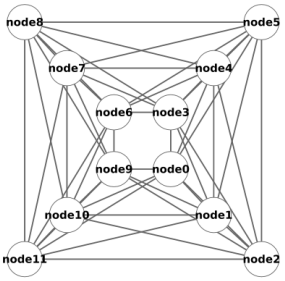


(a) Toy example of the fundamental assumption of a partition algorithm.

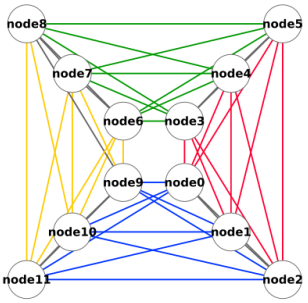


(b) How Infomap sees the network structure using random walks.

Fig. 2: Example of the partition approach to community discovery.



(a) Toy example of the fundamental assumption of a overlap algorithm.



(b) How HLC sees the network structure using node Jaccard.

Fig. 3: Example of the overlap approach to community discovery.

another on any given step and p_i is the fraction of within-community movements that occur in community i plus the probability of exiting module i. Accordingly, $H(Q)$ is the the entropy of clusters names, or city names in our intuition presented before, and $H(P_i)$ the entropy of movements within cluster i, the street names in our example, including the exit from it. Since trying any possible partition in order to minimize $L(M)$ is inefficient and intractable, the algorithm uses a deterministic greedy search and then refines the results with a simulated annealing approach.

B. Overlapping Approach

The overlap class of algorithms rejects the fundamental assumption of the partition approach. Here, nodes are allowed to be in multiple communities, therefore they are densely connected also to nodes that are not part of the community, removing the sparser areas of the network outside the community. A simplified example representing this concept is depicted in Figure 3(a). Here, nodes are grouped in cliques of 6 nodes, connected the one with the other by cliques of three nodes. So the nodes $\{0, 1, 2\}$ form a 3-clique with each other and two separated 6-cliques with nodes $\{3, 4, 5\}$ and $\{9, 10, 11\}$. Similar structures are generated by all other 3-node groups on the diagonals.

Even in this simple case there is no reasonable partition of the graph. There is no reason for which we should prefer clique $\{0, 1, 2, 3, 4, 5\}$ over clique $\{0, 1, 2, 9, 10, 11\}$, and we cannot merge them either, ignoring the fact that the other nodes are densely connected to them too. That is when an overlapping approach like HLC [6] proves its usefulness.

HLC assumes that communities should group together edges, not nodes. The relationship is part of a community and the node is part of all the communities its relationships are part of. In the case of a social network, a person knows other people for one main reason (work together, study together, spend together the free time, and so on) and therefore she is part of a different community for each "relationship environment". As a consequence, these communities overlap. Figure 3(b) depicts an example of HLC output for the graph presented in 3(a): each link is colored according to the link cluster it belongs to, and therefore we obtain as communities both $\{0, 1, 2, 3, 4, 5\}$ and $\{0, 1, 2, 9, 10, 11\}$.

For an undirected, unweighted network, we denote the set of node i and its neighbors as $n_+(i)$. HLC considers only link pairs that share a node, under the assumption that they are more similar than disconnected pairs. The similarity S between links e_{ik} and e_{jk} in the set E of all links in the network is computed as:

$$S(e_{ik}, e_{jk}) = \frac{n_+(i) \cap n_+(j)}{n_+(i) \cup n_+(j)}.$$

Shared node k does not appear in S because it provides no additional information and introduces bias. This is basically the Jaccard index of the set of nodes one step away from edges e_{ik} and e_{jk}. HLC then builds a link dendrogram from the presented equation (ties in S are agglomerated simultaneously). The dendrogram is cut at a S threshold that maximizes a quality function called "partition density". For each community, the partition density is defined as:

$$D_c = \frac{m_c - (n_c - 1)}{n_c(n_c - 1)/2 - (n_c - 1)},$$

where m_c is the number of links in the community c and n_c is the number of induced nodes in the community ($n_c = \bigcup_{e_{ij} \in c} \{i, j\}$). The overall partition density of a given set of link partition is the average of all partition densities, normalized over the total number of edges in the network:

$$D = \frac{2}{|E|} \sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)}.$$

V. Case Study

We now take a look at the characteristics of the results provided both by the partition and by the overlap approach. We consider the following list of characteristics:

- The distribution of community size, to understand if one method privileges larger or smaller communities, in Section V-A;

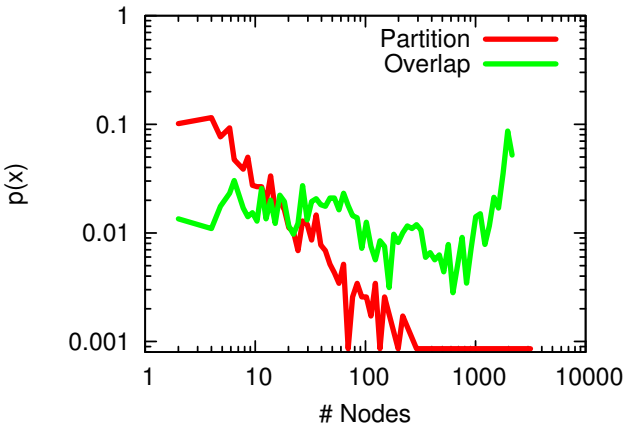


Fig. 4: The log binned distribution of the number of nodes per community, both for the partition and the overlapping approach.

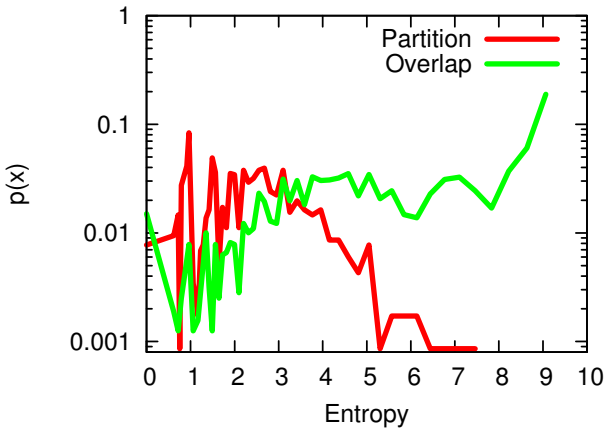


Fig. 5: The log binned distribution of the entropy per community, both for the partition and the overlapping approach.

- The community entropy w.r.t. the marketing classification, i.e. if the results of an algorithm are overlapping with the known product classes, in Section V-B;
- Community extracts, to provide examples of the typical communities returned by an algorithm, in Section V-C.

We then put together the discovered differences of community results in Section V-D.

A. Community Size

We start by providing a basic information about a community coverage: the distribution of the community size, i.e. the number of nodes per community. The distribution is depicted in Figure 4. On the x axis we report the number of nodes and on the y axis the probability that a community contains the given number of nodes, both for Infomap (red line) and HLC (green line). The distribution is log binned, i.e. each x axis value is grouped in bins of increasing size.

As we can see, the two distributions have different asymptotic behavior. Infomap provides a community size distribution that resembles a power-law, with more than 30% communities containing 4 nodes or less, and one community containing more than 3000 nodes. On the other hand, HLC communities have a very different size distribution: just above 3% of communities have 4 nodes or less, and 10% of them have around 2000 nodes. So the first difference between the two approaches can be summarized as: "The overlap approach returns larger communities than the partition approach".

B. Community Entropy

We now want to describe what is the actual content of these communities. In particular, we are interested in how homogeneous the communities are w.r.t. the marketing classification of the supermarket. In practice, we want to know if in a given community we grouped the products that belong to the same marketing classification. For the marketing classification, we use the Segment level, as presented in Section III. A good measure to do this is to calculate their information entropy. The information entropy is formally defined as the average unpredictability in a random variable, which is equivalent to its information content. In our case, a community c of $|c|$ nodes is viewed as $|c|$ outcomes of a random selection of a marketing classification. The possible outcomes of the extraction are $|M|$, the number of marketing classifications. The information entropy of a community $c \in C$ is then calculated as:

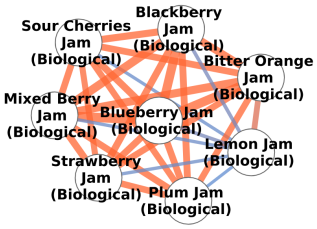
$$H(c) = -\sum_{m \in M} p(c_m) \log_2 p(c_m),$$

where $p(c_m)$ is the number of nodes in the community c that belongs to the marketing category m, over the total number of nodes inside community c. The average entropy value, calculated for all communities in C that is the community set, for Infomap is 1.8302, while the average for HLC is equal to 5.66305. The average entropy could not be an accurate information, as it may be driven by extreme values. For this reason, we depict in Figure 5 the probability (y axis) that a given community takes a given entropy value (x axis) for the communities extracted by Infomap (red line) and by HLC (green line). Again, the distributions are log binned.

Also in this case, the entropy distribution for Infomap and HLC look different. Most communities returned by Infomap have entropy lower than 3, and the number of communities with entropy larger than 6 is not significant. On the other hand, the majority of HLC communities have entropy larger than 3, with 20% of the communities having an entropy around 9.

One could think that the higher entropy of the HLC communities is due exclusively to the fact that HLC communities are larger on average. To disprove the objection, for each community of size $|c|$ we normalize the obtained entropy value over the description length required to code a random community, that is $\log_2 |c|$. We then sum up all the normalized values and take the community average, as:

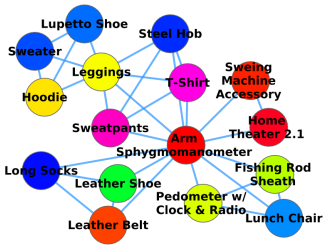
$$\bar{H}(C) = \frac{1}{|C|} \sum_{c \in C} \frac{H(c)}{\log_2 |c|}.$$



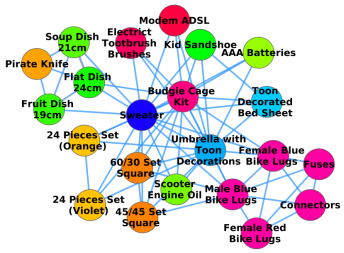
(a) Community #37.

(b) Community #80.

Fig. 6: Extracts from the non-overlapping communities.



(a) Community #239.



(b) Community #590.

Fig. 7: Extracts from the overlapping communities.

If $\bar{H}(C) = 1$, then, on average, $\forall c$ it holds $H(c) = \log_2 |c|$, i.e. the distribution of marketing classifications in the community is practically random. If $\bar{H}(C) = 1$, then for each community c we have $H(c) > \log_2 |c|$, then the communities separate products of the same marketing category even if it would be expected to find them in the same community. If $\bar{H}(C) < 1$, then on average we find products of the same marketing category in the same community.

The lower the $\bar{H}(C)$ value, the more homogeneous the communities are on the marketing classification, independently on their number of nodes. We found that in Infomap $\bar{H}(C) = 0.60180796763$, while for HLC $\bar{H}(C) = 0.814926005465$. We can conclude that HLC communities have a 20% higher entropy than Infomap, independently on community size. So the second difference between the two approaches can be summarized as: "The overlap approach returns communities that contains more diverse typologies of products than the partition approach".

C. Community Extracts

The aim of this section is to provide some concrete instances of the findings described in the previous subsections. We provide two examples of small communities extracted using the Infomap partition method and two examples of communities extracted with the HLC overlap community detection.

Figures 6(a) and 6(b) are the two extracts from the Infomap community partition. Given that most Infomap communities are small (see Figure 4) it is easy to find representative communities of limited size. In this case, we limit ourselves to communities containing 8 nodes. These two communities are identified as community #37 and #80 respectively.

In Figures 6(a) and 6(b), the node color refers to the node marketing Segment (see Section III). Colors are not consistent across figures, i.e. even if nodes from different figures have the same color it does not mean they are in the same segment. Edge width and color are proportional to edge weight, that is the number of times the two products were bought in the same shopping session. In Section III we refer to this quantity as "Absolute Support" of the pair and the minimum value is equal to 10, i.e. each connected pair of products in the community has been sold at least 10 times. In the edge color map, orange indicates high weight, blue indicates low weight.

By inspecting communities #37 and #80 we can observe the following characteristics:

- The communities are very dense: in fact, they are cliques, where every node is connected to every other node, meaning that different customers have bought all possible combinations of these products at least once;
- Most links have high weight, meaning that the amount of customers buying these products in the same shopping session is high;
- All products in the communities are part of a very homogeneous class of products: in community #37 we have only biological jams, while in community #80 we have only liquid yogurts.

We now turn to examine communities extracted with the overlap HLC approach. They are depicted in Figures 7(a) and 7(b) and they are identified with IDs #239 and #590. Again, the edge width and color is representative of edge weight in the same scale of Figures 6(a) and 6(b), while node color indicates the marketing segment and it is not consistent across figures, due to the high amount of segments present in each community. We had to choose communities with a larger number of nodes, given the relative scarcity of small communities returned by HLC (see Figure 4).

By inspecting communities #239 and #590 we can observe the opposite characteristics we observed for the Infomap communities:

- The communities are dense but they are not cliques: they are rather cliques joint together by some products (for example, the arm sphygmomanometer plays a central role in community #239);
- Almost all links have low weight, meaning that the amount of customers buying these products in the same shopping session is low;
- Almost all products in the communities are part of a different marketing segment.

So we can summarize the results of this section by saying that: "Examples shows that characteristics and topology of communities returned by the overlap approach are very different from the results of the partition approach".

D. Community Interpretation

In this section we wrap up the results we presented in the previous sections, providing a tentative explanations that takes into account all of them. The conclusions of each section were:

- The overlap approach returns larger communities than the partition approach;
- The overlap approach returns communities containing more diverse typologies of products than the partition approach;
- Examples shows that characteristics and topology of communities returned by the overlap approach are very different from the results of the partition approach.

Our explanation is then that the overlap approach mostly reflect customer behaviors and possible expansions of them, while the partition approach returns a refined marketing classification. We support our explanation by noticing that customers usually buy products for different marketing segments because they have to satisfy different needs, this also implies that a community grouping together a "customer profile" should be larger and more diverse on marketing classifications. Being less dense, overlap communities also put together products that some customers bought together and some others, who bought similar products, did not buy together, identifying possible customized product suggestions to the marketing department.

On the other hand, the partition approach is more homogeneous on the existing marketing classification, but the disagreement points may be interesting to explore to refine it. The small communities suggest a fine grained marketing description and the high density and edge weight of them implies that the products have really something to do with each other. It is worthwhile to notice that Infomap can also be used with a hierarchical approach, by merging related communities at different levels. In this way, it is possible to reconstruct a full marketing hierarchy. Also HLC by nature returns a hierarchy, but of a different kind: it is a hierarchy of extended customer profiles, with different, but nevertheless useful, classification of customers' behaviors and sub-behaviors.

VI. Conclusion

In this paper, we investigated the different application scenarios that one can tackle with community discovery using different community definitions. We focused on networks of products co-purchased in a supermarket. In particular, we have showed that there is not a quantitative difference in how good or bad are the results obtained by searching for a disjoint community partition and the results obtained from an overlapping coverage search. There is rather a qualitative difference, i.e. different problem definitions. A partition approach has proven to be useful as an approach to a marketing product classification. An overlap approach, instead, can shed some light over a novel technique for customer profiling.

The present work can be extended along several other lines of research. First and foremost, the distinction between partition and overlap based approaches is just one of the many in the field of community discovery. Some review works [3] have come as far as to identifying more than seven macro definitions of communities in complex networks. It is possible that each of these definitions is going to provide answers for additional problem definitions. As a second point, the empirical study about the different practical applications of alternative methods of community discovery can be separated from the application scenario we considered in this paper. Instead of focusing on product networks, we can consider many different typologies of networks, covering a wider set of human activities and natural phenomena.

Acknowledgments

We thank the supermarket company Coop and Walter Fabbri for sharing the data with us and allowing us to analyze and to publish the results. This work has been partially supported by the European Commission under the FET-Open Project n. FP7-ICT-270833, DATA SIM.

References

[1] D. Pennacchioli, M. Coscia, S. Rinzivillo, F. Giannotti, and D. Pedreschi, "Explaining the product range effect in purchase data," in *Big Data*, 2013.

[2] T. Raeder and N. V. Chawla, "Market basket analysis with networks," *Social network analysis and mining*, vol. 1, no. 2, pp. 97–113, 2011.

[3] M. Coscia, F. Giannotti, and D. Pedreschi, "A classification for community discovery methods in complex networks," *Statistical Analysis and Data Mining*, vol. 4, no. 5, pp. 512–546, 2011.

[4] S. Fortunato and C. Castellano, "Community structure in graphs," in *Computational Complexity*. Springer, 2012, pp. 490–512.

[5] M. Rosvall and C. T. Bergstrom, "Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems," *PloS One*, vol. 6, no. 4, p. e18209, 2011.

[6] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.

[7] S. Hill, F. Provost, and C. Volinsky, "Network-based marketing: Identifying likely adopters via consumer networks," *Statistical Science*, vol. 22, no. 2, pp. 256–275, 2006.

[8] D. Pennacchioli, M. Coscia, F. Giannotti, and D. Pedreschi, "Calculating product and customer sophistication on a large transactional dataset," Technical Report, 2013.

[9] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, P. Buneman and S. Jajodia, Eds., FebJun–FebAug˜ 1993, pp. 207–216. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.40.6984

[10] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*. ACM, 2012, p. 3.

[11] E. Ferrara, "Community structure discovery in facebook," *International Journal of Social Network Mining*, vol. 1, no. 1, pp. 67–90, 2012.

[12] D. Jin, D. Liu, B. Yang, and J. Liu, "Fast complex network clustering algorithm using agents," in *Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on*. IEEE, 2009, pp. 615–619.

[13] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E*, vol. 76, no. 3, pp. 036 106+, Sep. 2007. [Online]. Available: http://dx.doi.org/10.1103/physreve.76.036106

[14] M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, "Demon: a local-first discovery method for overlapping communities," in *KDD*, 2012, pp. 615–623.

[15] J. Xie and B. K. Szymanski, "Towards linear time overlapping community detection in social networks," Feb. 2012. [Online]. Available: http://arxiv.org/abs/1202.2465

[16] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New Journal of Physics*, vol. 11, no. 3, p. 033015, 2009.