Journal of
*Theoretical and Applied
Electronic Commerce Research*

Article

Discovering Customer Purchase Patterns in Product Communities: An Empirical Study on Co-Purchase Behavior in an Online Marketplace

Kenan Kafkas [1,*], Ziya Nazım Perdahçı [2] and Mehmet Nafiz Aydın [1]

[1] Department of Management Information Systems, Kadir Has University, Istanbul 34083, Turkey; mehmet.aydin@khas.edu.tr

[2] Informatics Department, Mimar Sinan Fine Arts University, Istanbul 34380, Turkey; nz.perdahci@msgsu.edu.tr

* Correspondence: kenankafkas@gmail.com

Abstract: Marketplace platforms gather and store data on each activity of their users to analyze their customer purchase behavior helping to improve marketing activities such as product placement, cross-selling, or customer retention. Market basket analysis (MBA) has remained a valuable data mining technique for decades for marketers and researchers. It discovers the relationship between two products that are frequently purchased together using association rules. One of the issues with this method is its strict focus on binary relationships, which prevents it from examining the product relationships from a broader perspective. The researchers presented several methods to address this issue by building a network of products (co-purchase networks) and analyzing them with network analysis techniques for purposes such as product recommendation and customer segmentation. This research aims at segmenting products based on customers' purchase patterns. We discover the patterns using the Stochastic Block Modeling (SBM) community detection technique. This statistically principled method groups the products into communities based on their connection patterns. Examining the discovered communities, we segment the products and label them according to their roles in the network by calculating the network characteristics. The SBM results showed that the network exhibits a community structure having a total of 309 product communities, 17 of which have high betweenness values indicating that the member products play a bridge role in the network. Additionally, the algorithm discovers communities enclosing products with high eigenvector centralities signaling that they are a focal point in the network topology. In terms of business implications, segmenting products according to their role in the system helps managers with their marketing efforts for cross-selling, product placement, and product recommendation.

Keywords: market basket analysis; co-purchase network; community detection; SBM; product segmentation

check for updates

Citation: Kafkas, K.; Perdahçı, Z.N.; Aydın, M.N. Discovering Customer Purchase Patterns in Product Communities: An Empirical Study on Co-Purchase Behavior in an Online Marketplace. *J. Theor. Appl. Electron. Commer. Res.* **2021**, *16*, 2965–2980. https://doi.org/10.3390/jtaer16070162

Academic Editor: María Teresa Ballestar

Received: 24 August 2021
Accepted: 23 October 2021
Published: 28 October 2021

1. Introduction

An online marketplace is a platform where multiple third-party companies provide services or commodities. The platform is essentially responsible for delivering the services that facilitate transactions between its users, namely, the buyers and sellers. These popular online platforms, such as Amazon or eBay, offer buyers the opportunity to make purchases on the same platform without leaving the site or application. These marketplaces gather and store several types of data about their users, one of which is the transaction data used to analyze the customer purchase behavior helping to improve marketing activities.

Market basket analysis (MBA) is a frequently used data mining method for such purposes. It discovers the relationship between two products that are frequently purchased together using a technique called association rules [1,2]. Although there have been significant contributions from an MBA point of view, there is a limitation on the method's effectiveness [3] because of its focus on only the binary relationship between two products. Researchers frequently apply the network science approach to established research fields to overcome its limitations [4]. To address MBA's binary relationship issue, researchers presented a network analysis [5–7] approach that helps to analyze not just the relationship between two products but also a whole network of relationships among all products in the system.

In this research, we empirically study the transaction data of an online marketplace platform. We build a co-purchase network by connecting products if they are purchased by the same customer. We then analyze the network by discovering the product communities based on the customers' co-purchase patterns. Certain products play a key role in the network by connecting otherwise isolated communities. Some products play a different role in the system by connecting highly connected products. We calculate two key centrality measures to discover such important products: eigenvector and betweenness centralities. Additionally, we include the total spending data to distinguish products monetarily. Despite various studies to discover the purchase patterns with a network approach, one of the concerns includes issues with community detection methods such as taking a heuristic path or a tendency to overfit the data. In this research, we employ the stochastic block modeling (SBM) method from the repertoire of community detection algorithms, a principled statistical inference method that groups the products based solely on their connections to discover latent product communities in the network.

This paper aims to segment the products by detecting the similarities in customers' co-purchase patterns by extending the MBA. The main focus of this study is to determine the roles of the products in the network and utilize the findings for improving marketing activities such as product placement, cross-selling, or customer retention. Despite its many alternatives, SBM is a statistically principled method, making its results domain independent and less error prone. Thus, it is a scientific technology suitable for decision support systems for any kind of electronic commerce.

The rest of the paper is organized as follows. Section 2 describes the related works and literature review, and Section 3 discusses the theoretical background. The proposed method framework is presented in Section 4. Furthermore, results are presented in Section 5, and finally, the paper closes with a discussion and conclusion in Section 6.

2. Related Works

Market basket analysis (MBA) is considered the most common way to understand co-purchase behavior both in the industry and in academia [8,9]. Agrawal et al. [1] describe MBA as follows: for products X and Y, if the same customer purchased Y while buying X, there is an "association rule" between X and Y, indicating a potential purchase pattern. Liao et al. [10] incorporate k-means clustering algorithm into the MBA to perform product segmentation. Their work presents managerial implications such as finding candidates for product bundling and new products to enter the market. In a recent study, Puka and Jedrusik [11] similarly use MBA and extend the association rules by combining it with the complementarity concept called basket complementarity. However, the methods based on association rules focus on only the relationship between two products. Ding et al. [7] point out the lack of network understanding "However, researchers have noticed that there are still many deficiencies in the market basket analysis, which deteriorates its effectiveness as a market analysis approach. One outstanding issue with market basket analysis stems from its focus solely on the 'association rules' between two products; in the real business context, however, there may be links between any products which form a group. Retailers are no longer satisfied by the analysis of binary relationships among products. They seek a whole picture of inter-product relationships, as traditional market basket analysis "is often difficult to isolate interesting relationships" [12]. Ding et al. [7] argue that "products that are not often purchased together may be used in similar scenarios, which are often overlooked or an implicit factor in the market basket analysis".

Many researchers applied the network analysis idea to go beyond this binary approach and understand the entire set of relationships in the system. Table 1 illustrates a comparison between nine representative studies that employ a community detection method on co-purchase data. In e-commerce literature, network understanding is generally introduced as an extension of MBA. To achieve that, researchers add basic network measures such as centrality to the traditional MBA [6]. Many researchers go further and add community detection to the research [12], which is an effort to split the network into groups based on the density of their connections. In addition, it is an established notion in network science that there is no single detection method that fits all situations summarized as "No Free Lunch Theory" [13,14], meaning that one should utilize the most appropriate detection method for the existing system. Modularity maximization is a heuristic method commonly used to detect communities in academia that tends to overfit the data and [15] has a resolution limit that prevents it from detecting small communities in large networks [16]. Nevertheless, it is the most common method that researchers employ.

Table 1. A summary of studies from literature in terms of four criteria involving co-purchase networks.

Researchers	Research Focus	Analysis Method	Attribute Used for Segmentation	Community Detection Methods/Heuristics Used
Clauset et al., 2004	Product Recommendation	Network partitioning	Not used	Modularity Maximization
Huang et al., 2007	Product Recommendation	Network partitioning	Not used	Random Graph Modeling
Raeder and Chawla, 2010	Discover relationship between products using network approach	Extending MBA with network approach	A novel metric "utility of community"	Modularity Maximization
Kim et al., 2012	Compare MBA networks with co-purchase networks	Extending MBA with network approach using a time limit	Degree centrality	K-Nearest Neighbors
Videla-Cavieres and Rios, 2014	Discover relationship between products more efficiently	Extending MBA with network approach	Not used	Modularity Maximization
Faridizadeh et al., 2018	Product Recommendation	Extending MBA with network approach	Degree centrality, density	Modularity Maximization
Ding et al., 2018	Discover relationship between products using network approach	Extending MBA with network approach	Betweenness centrality	Hierarchical SBM/K-Core Decomposition
Gabardo et al., 2019	Product Recommendation	Extending MBA with network approach	Not used	Modularity Maximization for overlapping communities
Chattopadhyay et al., 2020	Product Recommendation	Extending MBA with network approach	Node similarity	A method based on node similarity (nodality)
This research	Product segmentation	Extending MBA with network approach	Betweenness, eigenvector centralities and Monetary attribute	Degree-corrected Hierarchical Weighted SBM

Co-purchase networks generally have been studied to extend the standard MBA or to enhance recommendation systems. A considerable amount of literature has been published utilizing community detection methods to identify similar groups in the network [6,17–19]. However, much of the research has either applied problematic detection methods such as modularity maximization [20] or focused on basic centrality measures or clustering behaviors to analyze the network [6,19,21]. The study of Raeder and Chawla [12] is one of the early examples of using network approach to extend MBA. They detect communities using modularity maximization and propose a measure named utility of community which is a value derived from the number of edges to determine the role of the products in the network. However, to reduce the data set, they utilize a questionable method by "pruning" the network, which compromises the integrity of the network structure. Kim et al. [6] take a similar dataset of transaction data from a department store and model two different co-purchase networks. One connects two products if they appeared in the same ticket, and the other connects two products regardless of the time of purchase. They run the k-nearest neighbors algorithm to discover the communities and use degree centrality to detect the importance of the products. Our method involves eigenvector centrality an advanced version of degree centrality that not only reflects the number of connections of a product but also the number of connections of its neighbors. Videla-Cavieres and Rios [5] aim to extend MBA by utilizing network analysis techniques proposing a method to analyze large networks containing more than a hundred thousand nodes. As in [12] their method involves filtering edges to reduce the network to manageable sizes; however, removing edges of a network might compromise the underlying network structure. The present study covers the entire transaction data. Moreover, contrary to many studies [5,6] our method includes the co-purchases even if they take place only once.

Unlike the methods used in these studies, the SBM community detection method offers a probabilistic model, a principled statistical inference method [22] that discovers communities based on connection patterns of the nodes. We present its theoretical background in the next section. In the co-purchase network context, connections represent customers' purchases; therefore, the SBM method groups the products based on their buyers' purchase patterns. The methods used in previous studies, such as modularity maximization and K-core decomposition lack such properties.

Only Ding et al. [7] employ SBM among the studies seen in Table 1. Additionally, they take a more holistic approach that analyzes the network both at a macro level (hierarchies of the products) and micro-level (brokerage role of the products.) Utilizing the recent advancements in the field, researchers use three different community detection methods, one of which is Hierarchical Stochastic Block Modelling [23]. This holistic approach extends the binary perspective of the existing MBA, which focuses on the relationship of only two products to the whole network structure. Not all studies on co-purchase networks focus on MBA. For example, Gabardo et al. [24] and Chattopadhyay et al. [25] contribute to the co-purchase network research to improve product recommendation by bringing novel community detection methods based on overlapping communities and node similarity concepts, respectively. This research utilizes degree-corrected, hierarchical, weighted SBM which is a statistically principled method to discover product communities and ranks the products based on their monetary, betweenness and eigenvector attribute afterwards.

There are various methods to achieve product segmentation. Artificial neural networks are a recent example. Wang et al. [26] use self-organizing map, an artificial neural network method to segment the products. Additionally, they incorporate recency, frequency, and monetary analysis into their research. Apart from co-purchase analysis, product segmentation can be performed based on demographic data. For instance, Lees et al. [27] present demographic product segmentation in financial services using attributes such as gender, age, and socio-economic status. However, by discovering the product groups based only on customer purchase behavior, the present study performs a behavioral product segmentation.

3. Theoretical Background

3.1. Stochastic Block Model Community Detection

Finding latent communities in complex networks is a challenging task. One promising method in this space is the stochastic block model, which falls into the statistical inference group among the community detection methods. It is developed by social scientists in the 1980′s [28] to generate random networks that contain inherent community structure. When run in reverse fashion, SBM is used to infer latent communities within a given network (Figure 1).

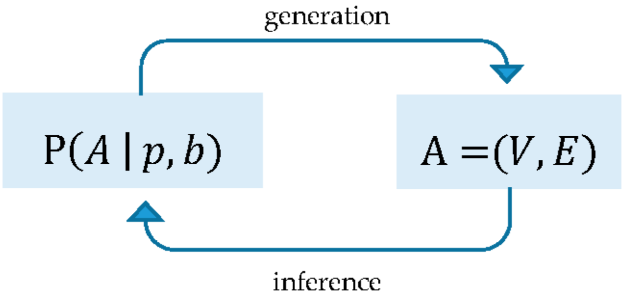


Figure 1. SBM generation and inference model. Generation: given probability distribution (p) of blocks b, draw network A. Inference: given network A (V vertices, E edges) choose p that makes A likely.

This relationship between generation and inference gives SBM a unique advantage against its alternatives, making it a benchmark community detection method. In this research, SBM is our choice of community detection method to discover product communities in the co-purchase network to reveal the hidden purchase behavior of the buyers.

3.1.1. Generative Aspect of SBM

For generating a random network that consists of desired blocks (groups, communities) one should provide the probability:

$$P(A|b)$$

where $A = \{A_{ij}\}$ is adjacency matrix that represents the network and b is a vector with $b_i \in \{1, \ldots, B\}$ entries that represent the building blocks of the network. Given the above information, SBM generates a network with Equations (1) and (2):

$$P(A|p, b) = \prod_{i<j} P_{b_i,b_j}^{A_{ij}} \left(1 - P_{b_i,b_j}\right)^{1-A_{ij}} \tag{1}$$

$$P_{rs} = e^{-\mu_{rs}} / \left(1 + e^{-\mu_{rs}}\right) \tag{2}$$

where P_{rs} is the probability of existence of an edge between two nodes from groups r and s.

3.1.2. Inference Aspect of SBM

For the inference side of SBM, instead of generating a network, the goal is to determine the probability of block b for a given network A.

$$P(b|A)$$

where acquiring this probability is called community detection in network science and it is performed by using Bayes' rule (Equation (3)) where $P(b|A)$ is the posterior distribution. This modeling approach makes this method a principled method rather than a heuristic one.

$$P(b|A) = \frac{P(A|b)\, P(b)}{P(A)} \tag{3}$$

3.2. SBM Types

There are several versions of SBM; we employ a combination of three of its versions in this study: degree corrected SBM [29], hierarchical SBM [23], and weighted SBM [30,31]. The standard SBM assumes that the probability of nodes connecting them to each other within a community is equal, which does not agree with real-world networks. This assumption makes the standard method sensitive to high degree nodes. Karrer and Newman [29] proposed a degree-corrected version of SBM to overcome this issue. Another issue in community detection is that on large networks, a resolution limit problem emerges, which prevents algorithms from detecting smaller but well-defined communities. The hierarchical SBM method addresses this issue by grouping communities as nested layers in a tree structure. As for the weighted SBM, it incorporates the edge weights into the algorithm and tries to fit the distribution of the weights to the target community. The edge weights are values that indicate the strength of connections between nodes in the network. In our case, sum of the money spent for both products at each end of an edge is used as edge weight. In other words, the total amount of money spent on products of a co-purchase pair will be the weight attribute of the weighted SBM. We will be using a combination of all three versions. Therefore, our method can be called degree corrected, weighted, hierarchical SBM (Equation (4)).

$$P(b|A,\, x) = \frac{P(x|A,\, b)\ P(A|b)P(b)}{P(A,\, x)} \tag{4}$$

where x is a model for weights between blocks, depending on the type of data, the algorithm allows us to use weight models such as exponential, normal, and binomial.

Furthermore, this largest component consists of hundreds of thousands of products belonging to several communities. We apply a community detection algorithm to this network, which then assigns the products into distinct communities based exclusively on the similarity of their connection patterns. In other words, products that fall into the same community exhibit a similar connectivity pattern. In this study, we exploit this similarity concept to segment the products. Further inspecting the structure of each community, we examine the products that belong to the same community in terms of their attributes that indicate their importance not only in their community but throughout the whole network.

Furthermore, we add the monetary aspect of the products as well as the size of the community. Finally, we use the resulting attribute composition to label the community and segment its member products that may answer the questions: Is there a specific product that plays a unique role in the community? Answers to this question may help managers make decisions on product segmentation, placement, and promotions.

3.3. Centrality Measures

In addition to discovering groups of nodes in the network, finding out the role of individual nodes throughout the entire network extends analysis. A set of measures called centrality measures quantifies how central a node is in the network. In this study, we group similar products then look at the two basic centrality measures of the group members to evaluate both the products, and the communities. The first one is betweenness centrality [32] that emphasizes the vertices which play a bridge role on the shortest paths from one vertex to another. Freeman introduced it to quantify how a person controls the information flow between other people. Consequently, high betweenness score nodes imply a strategic role as gatekeepers in the network.

The second measure we employed is the degree centrality. The most direct way to measure how central a vertex in a network is to count the number of connections to other vertices. However, having many connections to less connected vertices is not the same as having few connections to highly connected vertices. Eigenvector centrality algorithm [33] captures this nuance quantifying the centrality of a vertex accordingly.

4. Materials and Methods

The proposed framework for product segmentation is presented in Figure 2. It consists of three main steps: data preparation, network construction and community detection.

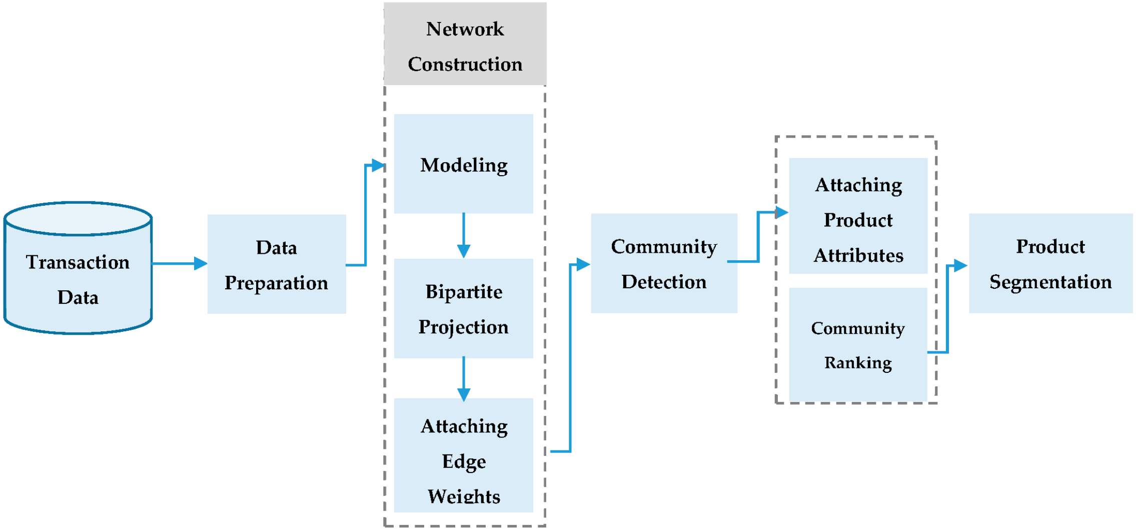


Figure 2. The method diagram of the research from raw transaction data to product segmentation.

4.1. Constructing the Network

4.1.1. Data Preparation

The raw data set contains nearly 1.5-million transactions obtained from one of the leading online marketplace platforms in Turkey where sellers offer a wide range of products. The transactions took place between 620,767 buyers and 7516 sellers involving 412,419 products. Time span of the transactions is three consecutive months. The data contains details of the transactions such as price amount, date and category information along with buyer attributes such as age and gender. However, we did not incorporate the demographic information in the present study. Among transactions, a small number of shipping fees shown as products had to be removed. In this study, we worked on a portion of the transactions spanning a two-week time frame in May 2015, which contains 228,026 transactions that took place between 139,885 unique buyers and 107,689 unique products.

4.1.2. Modeling

The first step is building a network from the data set. There are many ways to construct a network and it starts with deciding which entities in the data set will become the nodes and what will constitute the relationship between those entities (edges). Making this decision is called modeling the network. As online marketplace platforms facilitate transaction between buyers and sellers, the accumulated transaction data contains such entities as buyers, sellers and products which are all suitable candidates for being nodes in a network.

The edges in the network represent the relationship between chosen nodes which can be a purchase between a seller and a buyer or a message from a buyer to a seller. One of the frequently studied models is co-purchase networks which will be our focus in this research. Co-purchase here, implies that two products are purchased by the same buyer. Therefore, in a co-purchase network two products are connected to each other only if both are purchased by the same buyer or buyers. In online markets this type of relationship is typically referred to as “the customer who bought this item also bought this item” in product recommendation.

4.1.3. From Bipartite to Projection

To link two co-purchased products, we should first create a bipartite network where there are two distinct types of nodes: buyers, and products. We draw an edge between a buyer and a product in this model if the buyer has purchased the product. In bipartite networks two types of nodes never link among themselves, they only connect with the opposing type. Figure 3b shows a simple model of a product-buyer bipartite network along with two projections at both sides.

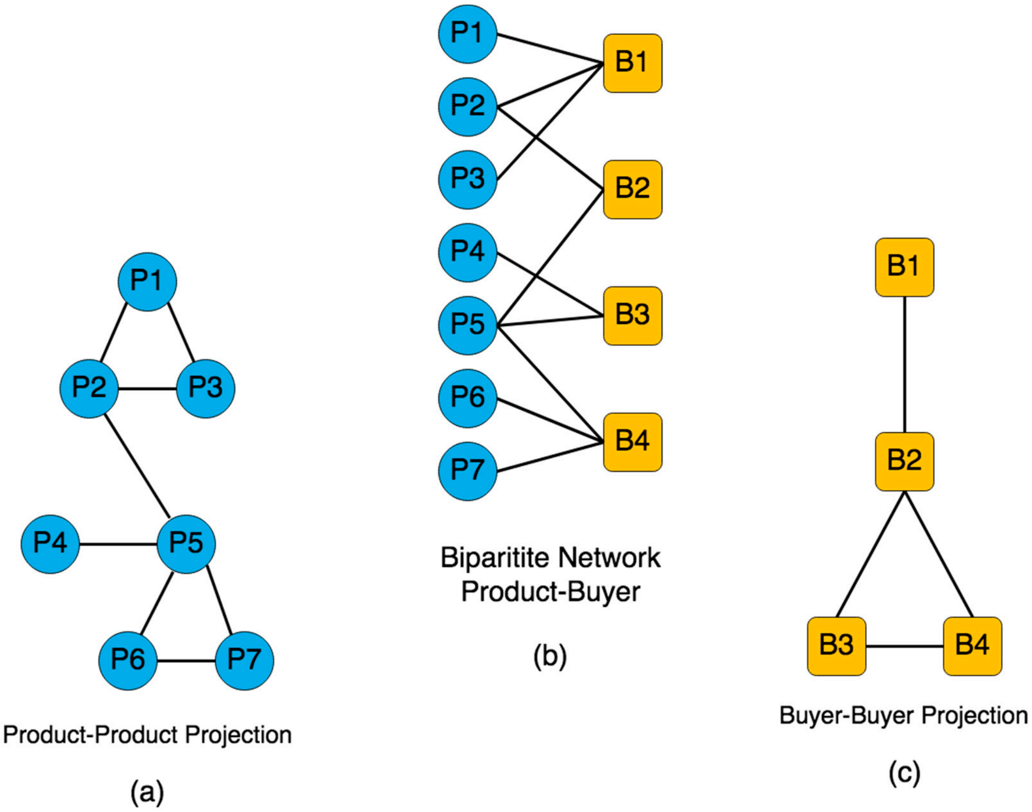


Figure 3. Bipartite network model and its two projections (**b**). Blue nodes are products and yellow nodes are buyers. Undirected projections: product-product (**a**) the co-purchase network, buyer-buyer (**c**).

To generate a co-purchase network, we split the bipartite network into two undirected subnetworks called projections. One of the projections will be buyer to buyer network, where an edge between two buyers indicates two buyers who bought the same product (Figure 3c). The other one will be the product-to-product projection, where an edge between two products means two products are bought by the same buyer or buyers. We discard the former one and work on the latter, the co-purchase network (Figure 3a). Following a similar approach, one can choose other options such as product-seller bipartite network which can be split in two projections: product-product and a seller-seller networks. However, we will keep the scope of this research limited to previous co-purchase network illustrated in Figure 3a.

4.1.4. Attaching Edge Weights

The product-product (co-purchase) network is undirected meaning the edges have no direction from one product to another and it is modeled in such a way that two products are connected only if they are purchased by the same buyer. However, several other buyers may also have bought the same two products together and such buyers most probably have varying attributes in terms of their platform value. Additionally, buyers are not the only actors in a marketplace platform, sellers also are an important part of the transaction. They have their own attributes that can contribute to the analysis of the complex system as well. We can assign such attributes to the network as node and edge attributes. Node attributes are attached to the products, and they indicate the value of the products e.g., price, category, number of transactions, etc. As for the buyers and sellers, the information indicating their value is attached to the connections between products. They are called the edge attributes (weights) which will play an important role in our analysis. Figure 4 is a simple model showing how the attributes are attached to the network on both nodes and edges. A list of possible information that can be used as node or edge attributes extracted from the transaction data is shown in Table 2. However, in this study we utilize only the monetary aspect which is the total amount of money spent for the co-purchase pairs (total spending), by aggregating total paid amounts of products at both ends of an edge. For instance, assuming two products P1 and P2 in Figure 4 are co-purchased by several buyers, we sum up the total paid amounts for both products and attach this value as an edge weight in the co-purchase network. Instead of total spending, a different study can be carried out using frequency of the purchases as the edge weights that can reflect differently on the research findings.

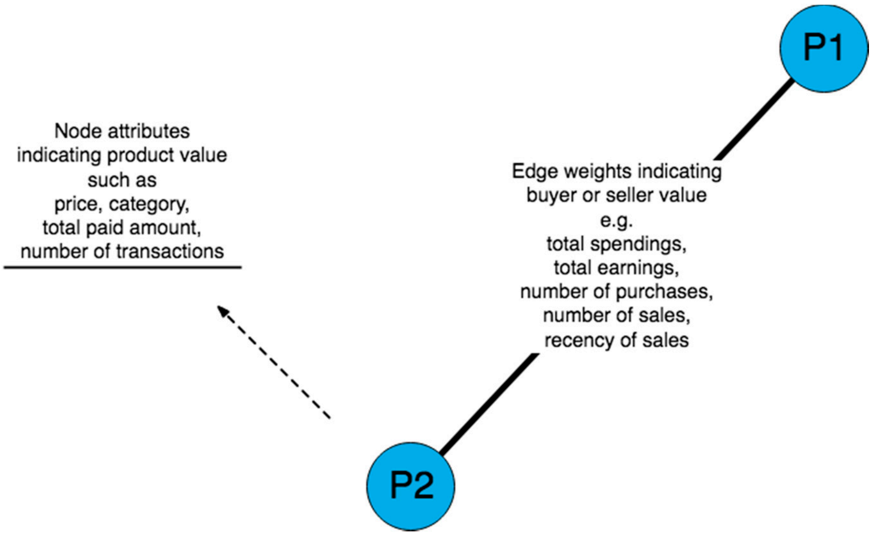


Figure 4. Co-purchase network model showing potential edge weights and node attributes. Only the amount of money spent is used in the study (monetary attribute).

Table 2. List of potential edge weights and node attributes that can be extracted from the transaction data.

Product (Node)	Buyer (Edge)	Seller (Edge)
Price Category Total paid amount Number of transactions	Frequency of purchases Recency of purchases Total spending Number of Purchases Age (sparse) Gender (Sparse) Subscription time	Frequency of sales Recency of sales Total earnings Number of sales

Following the network construction, we focus on discovering the product communities where the products are grouped together, signaling a similarity. In Figure 3a we can spot two communities at first glance (P1, P2, P3) and (P4, P5, P6, P7). Surely, we did not use an algorithm to detect those groups, we only performed a visual inspection. There are several community detection methods that can find the clustering of the nodes for us algorithmically. We will use a community detection algorithm chosen from a vast number of algorithms available where many of them detect different aspects of the network community structure depending on their 'community' definition.

4.2. Community Detection for Product Segmentation

For finding a good estimate for community detection a greedy algorithm based on merge-split Markov chain Monte Carlo (MCMC) is performed [34]. We performed several runs with varying numbers of Monte Carlo sweeps and iterations on one-, two- and four-week co-purchase networks. We then plotted the entropy for each iteration to track the minimization process to find the optimum iteration number and decided to run the algorithm with 10 sweeps for 200 iterations.

4.3. Attaching Product Attributes

Up to now, the nodes have no attributes other than their product ids. We calculate betweenness and eigenvector centrality scores of each product in the network. Additionally, monetary attribute of each product is attached to the network. Naturally, the attributes exhibit varying ranges of values for instance, betweenness score always ranges between 0 and 1, whereas monetary attribute may range from 0.5 to thousands of TRY. To be able to compare their values we calculate the rank of each value using fractional ranking method. Furthermore, we normalize their ranks as percentage values. For instance, a product with 92% betweenness score means that if all the betweenness attributes are ordered from 0 to 100 this product takes the highest 92nd place.

4.4. Ranking the Communities

After calculating the attributes of all products, we aim to find how those attributes are distributed in each community and use this composition to label them. For instance, to label a community of hundred products, one should determine the prominent characteristic in the community. If the community's mean betweenness attribute is significantly higher than other communities, we label this community as a high-betweenness community. If, however, the standard deviation of the attribute is not small then, one should not use this attribute to label the community. After labeling communities, we calculate the size of each community as an additional comparison parameter.

Using simple labels such as low, medium, and high instead of specifying the labels as percentages seems more suitable for comparison purposes. Moreover, the task of converting percentage values to three labels is not trivial, as the attributes may not be uniformly distributed over the communities to label mean percentages lower than 33% as low. To determine the transition thresholds of these levels, we plot the distribution of each attribute over the communities and look for appropriate percentage cutoff points. Due to the highly skewed distribution of community sizes, we split the sizes into three levels: small, medium, and large.

5. Results

The SBM algorithm discovered 309 product communities, and computation time took one hour 32 min to complete with 10 MCMC sweeps per iteration and 200 forced iterations in total. Attribute calculations took 32 min and calculating the buyer scores took an hour and 52 min using an Intel i5 CPU notebook with 12 GB of RAM.

Figure 5 shows the distribution of attribute percentages that helps us determine the cutoff thresholds, which we then use to label the community attributes as small, medium, or high as shown in Table 3.

Table 3. Cutoff thresholds for community attributes.

	Low (%)	**Medium (%)**	**High (%)**
Betweenness	0–55	55–80	80–100
Eigenvector	0–30	30–60	60–100
Monetary	0–20	20–60	60–100
	Small	Medium	Large
Size	0–20	20–350	350–6000

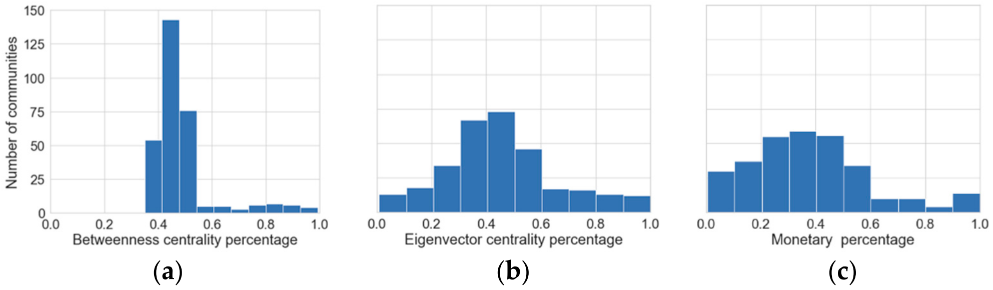


Figure 5. Mean percentage histogram of attributes vs. number of communities. Mean percentages of betweenness centrality (**a**), mean percentages of eigenvector centrality (**b**), mean percentages of monetary attribute (**c**).

Examining the betweenness attribute in Figure 5a, we observe that none of the communities have a mean percentage lower than 35%, and many communities lie between 35–55%. The rest are very low values, and they are almost equally distributed. The eigenvector centrality is close to a normal distribution (Figure 5b). As for the monetary attribute (Figure 5c), the range between 30% and 50% has the largest number of communities.

There is a community with 5543 products, another with 4683, and the following largest six communities contain between 1000 and 2000 products. We use the community size histogram to determine the cutoff thresholds for level labels; small, medium, large (Figure 6). Table 3 is a list of the cutoff points determined by examining their distributions.

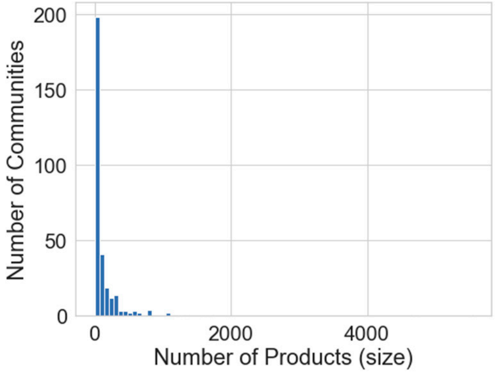


Figure 6. Histogram of community sizes. Communities.

Table 4 shows the breakdown of the number of community attributes which is determined by the thresholds given in Table 3. Seventeen communities have high-level betweenness attributes. In other words, the average betweenness centrality of those products is more than 80% compared to the rest of the communities.

Table 4. Number of communities for each category.

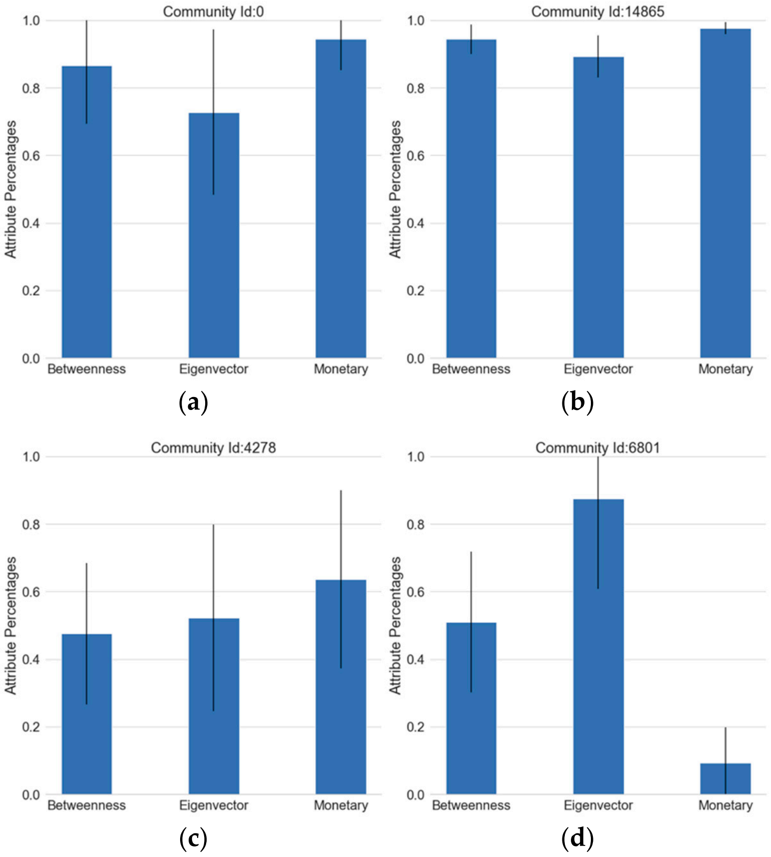
	Betweenness	**Eigenvector**	**Monetary**	**Size**
Low	273	64	66	66 (small)
Medium	19	184	205	218
High	17	61	38	25 (large)

Table 5 is the correlation matrix of the community attributes. The betweenness attribute highly correlates with the monetary attribute.

Table 5. Correlation matrix of the community attributes.

	Size	**Bet.**	**Eigen.**	**Mon.**
Size	1.000	0.033	0.078	0.191
Betweenness	0.033	1.000	0.446	0.646
Eigenvector	0.078	0.446	1.000	0.305
Monetary	0.191	0.646	0.305	1.000

Figure 7 shows four representative communities with various sizes and characteristics. The details of the communities in Figure 7 are shown in Table 6, listing the mean of the attribute percentages with their standard deviations and the mean percentage levels. To elaborate, the average of (normalized to 1) betweenness values (mean betweenness for short) of the products in community (a) is 0.87. The standard deviation of the normalized betweenness (S.D. for short) values of the products for the same community is 0.17. After ranking the mean betweenness of this community, its level is determined as “high” compared to the rest of the communities. The community in Figure 7a has high levels in all attributes, and there are 14 similar communities with various sizes. The community in Figure 7b exhibits similar values with one difference; namely that the standard deviations are much smaller. One of the largest communities in the network (Figure 7c) is an example of a monetary-dominant community. We assume an attribute as dominant if it has a high level while the other attributes are medium or low. Another example for dominant attributes is the community in Figure 7d having high eigenvector values on average. There are no betweenness dominant communities in the network. All high betweenness level communities show high levels in other attributes as well.



(**a**) (**b**)

(**c**) (**d**)

Figure 7. Mean attribute percentages and standard deviations of four selected communities. Community with high values in all attributes (**a**), small standard deviation values (**b**), high monetary attribute (**c**), high eigenvector attribute (**d**).

Table 6. Mean attribute values and standard deviations (S.D.) of four selected communities.

Community	Size	Betweenness			Eigenvector			Monetary		
		Mean	S.D.	Level	Mean	S.D.	Level	Mean	S.D.	Level
a	1175	0.87	0.17	High	0.73	0.24	High	0.94	0.09	High
b	10	0.94	0.04	High	0.89	0.06	High	0.98	0.02	High
c	4683	0.48	0.20	Low	0.52	0.28	Medium	0.64	0.26	High
d	243	0.50	0.20	Low	0.87	0.27	High	0.93	0.10	Low

A section of co-purchase network is shown in Figure 8 where there are two main product groups, groceries, and mobile phone accessories. Milk and phone case have high eigenvector centrality values whereas phone charger has high betweenness centrality value.

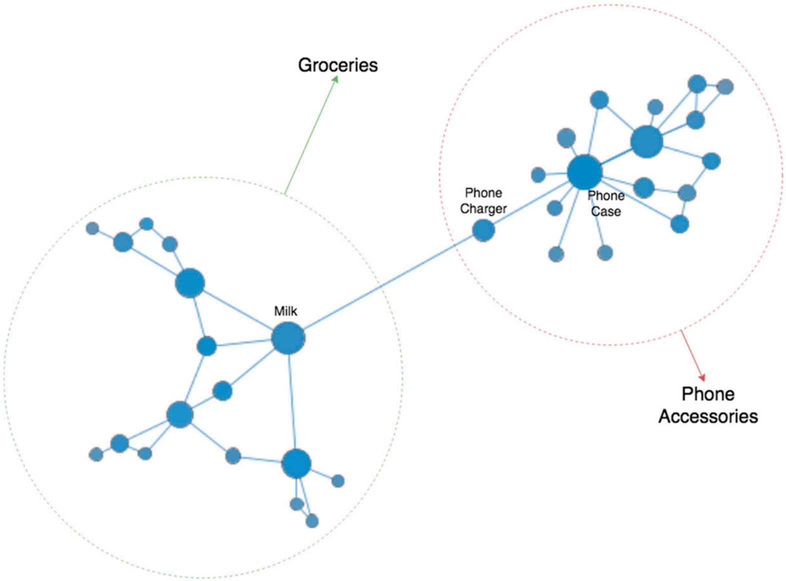


Figure 8. A subgraph of the co-purchase network showing two product groups with high centrality products (milk and phone charger).

6. Discussion

In this study, we apply a network approach to MBA, extending it with recent community detection algorithms and ranking the discovered communities based on the centrality attributes of their products. We build a product network based on co-purchase relationships and discover the product communities depending on the purchase behavior of their mutual buyers. Traditionally, market basket analysis is carried out on products purchased in one basket or one shopping trip. However, in the online marketplace context, a modern version of a shopping trip is physically almost effortless, enabling buyers to make purchases throughout the day or week, suggesting a new perspective on adapting the basket concept to current customer practices. To address this issue, we broadened the scope of the basket to two weeks.

Modularity maximization community detection method can find communities in a network even if there are no underlying communities in the network. One of the features of the SBM is that it can detect whether the network has a community structure or not. The results show that the co-purchase network has several communities. The algorithm discovers 309 product communities, eight of which contain more than one thousand products. The first thing we notice is that they contain medium or high-level monetary products, which is expected as we used this monetary attribute as the edge weight of the SBM algorithm. The correlation matrix in Table 5 supports this observation as we see that the highest correlating attribute with community size is the monetary attribute. Notice that although this correlation coefficient is the highest compared to other pairs (0.191), it is still a small value as the weight of SBM is not the only underlying factor in community detection.

The size of communities varies from a few products to thousands, as seen in Figure 6. To segment a product, we determine the dominant attribute of its community if one attribute is distinctly higher than the others. The first example is one of the largest communities with 1175 products which exhibits high levels in all attributes (Figure 7a). There are 14 such communities in the network. Following that, a small community with ten products also shows high levels in all attributes with minimal standard deviation values, increasing confidence in that measurement (Figure 7b).

A monetary dominant community (Figure 7c) indicates that high volumes of transactions took place for those products. However, their network centralities are not as significant as the others. They are high-volume products with low marketing value from a product recommendation perspective.

Faridizadeh et al. [19] use the degree centrality metric to assess the topological significance of the product in the network and argue that products with a high degree centrality are focal points in the network, indicating that they act as complementary products. Furthermore, those products can be recommended in cross-selling or up-selling activities. In this study, we find the communities that contain products with high eigenvector centrality values. The community in (Figure 7d) is an eigenvector-dominant community, which indicates that the products in this community are more topologically central. Eigenvector centrality indicates that a product is highly connected with other products. Unlike degree centrality, it shows neighboring products also have high connectivity. In a co-purchase network, this implies that they are star products frequently purchased with many other high degree products, making them good candidates for marketing efforts such as cross-selling, up-selling, and product placement.

Seventeen communities have high betweenness values. Except for two medium-level communities, all of which are high-level in eigenvector attributes as well. High-betweenness products connect two or more groups of products even if they are not highly connected. They serve as a gatekeeper between product groups. Ding et al. [7] argue that gatekeeper products interact with other product communities and adding that “They can be used as an introductory product of the community to stimulate the trial of new customers through the joint promotion with other product communities.” [7] In terms of business implications, their study concludes that segmenting products by their role in the network will help marketers to develop effective strategies in cross-marketing and new product launches. Using gatekeeper products, for instance, marketers can guide a customer interested in such a product towards a different group of products that are not directly related. In the network, we observe that phone chargers are frequently purchased with groceries. A phone charger can be recommended to a customer who purchases groceries. If the customer is interested in this recommendation, then a phone case or headphones recommendation follows. Thus, the phone charger plays the role of a gatekeeper between product groups guiding the customer from the groceries group to the phone accessories group.

7. Conclusions

This study discovered customers’ purchase patterns by examining product network communities using the stochastic block modeling (SBM), a principled method that uses Bayesian statistical inference. Being a probabilistic and generative model, SBM offers a superior solution to heuristics-based methods such as modularity maximization, which tends to overfit the data and suffers from discovering latent communities in large networks. This makes its results independent and less error prone. Thus, it is not only a scientific innovation but also a new scientific technology suitable for decision support systems for any kind of electronic commerce. This new scientific technology could be integrated into the existing decision support systems of market places online in a short period of time.

Segmenting the products based on customer purchase patterns and their role in the network helps marketing managers improve marketing activities such as product recommendation, product placement, cross-selling, or customer retention.

As a limitation for our research, the stochastic nature of the SBM causes the output to vary with only a few products being assigned to different communities at each run of the algorithm. In this study, we used the monetary attribute as the edge weights for the SBM. We observed its effects in the results as the algorithm tended towards putting monetarily similar products in the same communities. As future work, frequency or recency information can be selected to observe the results, or all potential edge weights can be used to determine which fits best to the data. We hope our research contributes to e-commerce literature by employing a principled approach.

Author Contributions: Conceptualization, K.K.; methodology, K.K.; software, K.K.; validation, Z.N.P. and M.N.A.; formal analysis, K.K.; writing—original draft preparation, K.K.; writing—review and editing, K.K., Z.N.P. and M.N.A.; visualization, K.K.; supervision, Z.N.P. and M.N.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Mimar Sinan Fine Arts University Scientific Research Projects Program (grant acronym: BAP, no: 2020-13).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Agrawal, R.; Imieliński, T.; Swami, A. Mining Association Rules between Sets of Items in Large Databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, USA, 25–28 May 1993; pp. 207–216.
2. Agrawal, R.; Srikant, R. *Fast Algorithms for Mining Association Rules*; Citeseer: Princeton, NJ, USA, 1994; Volume 1215, pp. 487–499.
3. Vindevogel, B.; Van den Poel, D.; Wets, G. Why promotion strategies based on market basket analysis do not work. *Expert Syst. Appl.* **2005**, *28*, 583–590. [CrossRef]
4. Esmaeili, L. Alireza hashemi golpayegani a novel method for discovering process based on the network analysis approach in the context of social commerce systems. *J. Theor. Appl. Electron. Commer. Res.* **2021**, *16*, 34–62. [CrossRef]
5. Videla-Cavieres, I.F.; Rios, S.A. Extending market basket analysis with graph mining techniques: A real case. *Expert Syst. Appl.* **2014**, *41*, 1928–1936. [CrossRef]
6. Kim, H.K.; Kim, J.K.; Chen, Q.Y. A product network analysis for extending the market basket analysis. *Expert Syst. Appl.* **2012**, *39*, 7403–7410. [CrossRef]
7. Ding, Z.; Hosoya, R.; Kamioka, T. Co-Purchase Analysis by Hierarchical Network Structure. PACIS 2018 Proceedings. 149. Yokohama, Japan, 2018. Available online: https://aisel.aisnet.org/pacis2018/149 (accessed on 27 September 2018).
8. Büchter, O.; Wirth, R. *Discovery of Association Rules over Ordinal Data: A New and Faster Algorithm and Its Application to Basket Analysis*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 36–47.
9. Woo, J. Market basket analysis algorithms with mapreduce. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2013**, *3*, 445–452. [CrossRef]
10. Liao, S.-H.; Chen, Y.-J.; Yang, H.-W. Mining customer knowledge for channel and product segmentation. *Appl. Artif. Intell.* **2013**, *27*, 635–655. [CrossRef]
11. Puka, R.; Jedrusik, S. A new measure of complementarity in market basket data. *J. Theor. Appl. Electron. Commer. Res.* **2021**, *16*, 670–681. [CrossRef]
12. Raeder, T.; Chawla, N.V. Modeling a Store's Product Space as a Social Network. In Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, Athens, Greece, 20–22 July 2009; pp. 164–169.
13. Peel, L.; Larremore, D.B.; Clauset, A. The ground truth about metadata and community detection in networks. *Sci. Adv.* **2017**, *3*, e1602548. [CrossRef]
14. McCarthy, A.D.; Chen, T.; Ebner, S. *An Exact No Free Lunch Theorem for Community Detection*; Springer: Cham, Switzerland, 2019; pp. 176–187.
15. Ghasemian, A.; Hosseinmardi, H.; Clauset, A. Evaluating overfit and underfit in models of network community structure. *IEEE Trans. Knowl. Data Eng.* **2019**, *32*, 1722–1735. [CrossRef]
16. Fortunato, S.; Barthelemy, M. Resolution limit in community detection. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 36–41. [CrossRef]
17. Ma'arif, M.R.; Mulyanto, A. Improving recommender system based on item's structural information in affinity network. *Proceeding Electr. Eng. Comput. Sci. Inform.* **2014**, *1*, 186–189. [CrossRef]
18. Oestreicher-Singer, G.; Libai, B.; Sivan, L.; Carmi, E.; Yassin, O. The network value of products. *J. Mark.* **2013**, *77*, 1–14. [CrossRef]

19. Faridizadeh, S.; Abdolvand, N.; Harandi, S.R. Market basket analysis using community detection approach: A real case. In *Applications of Data Management and Analysis*; Springer: Cham, Switzerland, 2018; pp. 177–198.
20. Newman, M.E. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8577–8582. [CrossRef]
21. Huang, Z.; Zeng, D.D.; Chen, H. Analyzing consumer-product graphs: Empirical findings and applications in recommender systems. *Manag. Sci.* **2007**, *53*, 1146–1164. [CrossRef]
22. Peixoto, T.P. Bayesian stochastic blockmodeling. In *Advances in Network Clustering and Blockmodeling*; 2019; pp. 289–332. Available online: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119483298.ch11 (accessed on 23 November 2019).
23. Peixoto, T.P. Hierarchical block structures and high-resolution model selection in large networks. *Phys. Rev. X* **2014**, *4*, 011047. [CrossRef]
24. Gabardo, A.; Berretta, R.; Moscato, P. Overlapping communities in co-purchasing and social interaction graphs: A memetic approach. In *Business and Consumer Analytics: New Ideas*; Springer: Cham, Switzerland, 2019; pp. 435–466.
25. Chattopadhyay, S.; Basu, T.; Das, A.K.; Ghosh, K.; Murthy, L.C. Towards effective discovery of natural communities in Complex networks and implications in E-commerce. *Electron. Commer. Res.* **2020**, *21*, 917–954. [CrossRef]
26. Wang, S.-C.; Hsu, H.-W.; Dai, C.-G.; Ho, C.-L.; Zhang, F.-Y. Use Product Segmentation to Enhance the Competitiveness of Enterprises in the IoT. In Proceedings of the IEEE 10th International Conference on Awareness Science and Technology (iCAST), Morioka, Japan, 23–25 October 2019; pp. 1–6.
27. Lees, G.; Winchester, M.; De Silva, S. Demographic Product Segmentation in Financial Services Products in Australia and New Zealand. *J. Financ. Serv. Mark.* **2016**, *21*, 240–250. [CrossRef]
28. Holland, P.W.; Laskey, K.B.; Leinhardt, S. Stochastic blockmodels: First steps. *Soc. Netw.* **1983**, *5*, 109–137. [CrossRef]
29. Karrer, B.; Newman, M.E. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **2011**, *83*, 016107. [CrossRef] [PubMed]
30. Aicher, C.; Jacobs, A.Z.; Clauset, A. Learning latent block structure in weighted networks. *J. Complex Netw.* **2015**, *3*, 221–248. [CrossRef]
31. Peixoto, T.P. Nonparametric weighted stochastic block models. *Phys. Rev. E* **2018**, *97*, 012306. [CrossRef] [PubMed]
32. Freeman, L.C. A set of measures of centrality based on betweenness. *Sociometry* **1977**, *40*, 35–41. [CrossRef]
33. Newman, M. The mathematics of networks. In *The New Palgrave Encyclopedia of Economics*, 2nd ed.; Blume, L., Durlauf, S.D., Eds.; Palgrave Macmillan: Basingstoke, UK, 2008.
34. Peixoto, T.P. Merge-split markov chain monte carlo for community detection. *Phys. Rev. E* **2020**, *102*, 012305. [CrossRef] [PubMed]