KADİR HAS UNIVERSITY
SCHOOL OF GRADUATE STUDIES
DEPARTMENT OF ADMINISTRATIVE SCIENCES

PRODUCT AND CUSTOMER SEGMENTATION BY PURCHASE BEHAVIOR IN E-COMMERCE PLATFORMS USING STOCHASTIC BLOCK MODEL

KENAN KAFKAS

PHD THESIS

ISTANBUL, JUNE, 2022

Kenan Kafkas Ph.D. Thesis 2022

PRODUCT AND CUSTOMER SEGMENTATION BY PURCHASE BEHAVIOR IN E-COMMERCE PLATFORMS USING STOCHASTIC BLOCK MODEL

KENAN KAFKAS

A thesis submitted to
the School of Graduate Studies of Kadir Has University
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in
Management Information Systems

Istanbul, June, 2022

APPROVAL

This thesis titled PRODUCT AND CUSTOMER SEGMENTATION BY PURCHASE BEHAVIOR IN E-COMMERCE PLATFORMS USING STOCHASTIC BLOCK MODEL submitted by KENAN KAFKAS, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Management Information Systems is approved by

Assoc. Prof. Dr. Mehmet Nafiz Aydın (Advisor) . . . .
Kadir Has University

Asst. Prof. Dr. Nazım Ziya Perdahçı (Co-Advisor) . . . .
Mimar Sinan Fine Arts University

Prof. Dr. Nimet Uray . . . .
Kadir Has University

Prof. Dr. Sona Mardikyan . . . .
Boğaziçi University

Assoc. Prof. Dr. Ferdi Sönmez . . . .
Fenerbahçe University

Asst. Prof. Dr. Deniz Eroğlu . . . .
Kadir Has University

I confirm that the signatures above belong to the aforementioned faculty members.

. . . .
Prof. Dr. Mehmet Timur Aydemir
Director of the School of Graduate Studies
Date of Approval: 15/06/22

DECLARATION ON RESEARCH ETHICS AND PUBLISHING METHODS

I, KENAN KAFKAS; hereby declare

- that this Ph.D. Thesis that I have submitted is entirely my own work and I have cited and referenced all material and results that are not my own in accordance with the rules;
- that this Ph.D. Thesis does not contain any material from any research submitted or accepted to obtain a degree or diploma at another educational institution;
- and that I commit and undertake to follow the “Kadir Has University Academic Codes and Conduct” prepared in accordance with the “Higher Education Council Codes of Conduct”.

In addition, I acknowledge that any claim of irregularity that may arise in relation to this work will result in a disciplinary action in accordance with university legislation.

KENAN KAFKAS

. . . .
15/06/22

To My Wife

ACKNOWLEDGEMENT

I would like to express my deep and sincere gratitude to my co-advisor, Asst. Prof. Nazım Ziya Perdahçı, who guided me with invaluable patience, friendship, and empathy throughout this thesis. I would also like to thank my research supervisor, Assoc. Prof. Dr. Mehmet Nafiz Aydın, for giving me the opportunity to do research and providing invaluable guidance throughout this research. It was a privilege and honor to work and study under their guidance.

I am extremely grateful to my wife for her love, care, and sacrifices. I also wish to extend my special thanks to my dear friend Assoc. Prof. Ahmet Salih Bıçakçı for his moral support.

Finally, my thanks go to all the people who have supported me in completing the research work.

Kenan Kafkas

PRODUCT AND CUSTOMER SEGMENTATION BY PURCHASE BEHAVIOR IN E-COMMERCE PLATFORMS USING STOCHASTIC BLOCK MODEL

ABSTRACT

To attract and maintain lucrative clientele, commercial internet platforms compete with a multitude of competitors by providing appropriate goods and services, employing a range of marketing methods to get a competitive edge utilizing their digital trace data. Techniques include a variety of marketing tactics, many of which are based on updated versions of conventional marketing strategies. Working out what consumers want and how to meet their needs is an ongoing task on these platforms. The literature is constantly being enhanced by new theoretical and practical applications. Customer purchase behavior leaves digital trace data in online platforms such as clickstream, transaction, or product review forms. This thesis proposes a model that presents a novel network approach to customer behavior analytics on online transaction data to perform product and customer segmentation. We seek answers to the following research questions: Can we understand the customer behavior and preferences through network analysis? If there are several purchase behavior types, what are the underlying patterns? Are there certain special products that play a special role in the network? To support decision-makers in their endeavor to improve marketing activities such as targeted advertising, increasing brand loyalty, attracting desired customers, and signaling more effective marketing messages. We utilize the Stochastic Block Model (SBM), which is a statistically principled community detection method on co-purchase networks to discover latent product communities, and we produce two different segmentation methods based on those communities. The outcome is a product and a customer segmentation which extends traditional data mining methods. We combine product based segmentation with Market Basket Analysis and customers segmentation with the RFM models. We implement our model on two empirical data sets. Lastly, we provide an executive summary for both examples.

Keywords: Customer Segmentation, Stochastic Block Model, Co-purchase Network, Community Detection, Diversity

STOKASTİK BLOK MODEL KULLANARAK E-TİCARET PLATFORMLARINDA SATIN ALMA DAVRANIŞINA GÖRE ÜRÜN VE MÜŞTERİ SEGMENTASYONU

ÖZET

Kazançlı müşteri kitlesini çekmek ve sürdürmek için, ticari internet platformları, dijital izleme verilerini kullanarak rekabet avantajı elde etmek için bir dizi pazarlama yöntemi kullanırlar, Uygun mal ve hizmetleri sağlayarak çok sayıda rakiple rekabet ederler. Kullanılan teknikler, çoğu geleneksel pazarlama stratejilerinin güncellenmiş versiyonlarına dayanan çeşitli pazarlama taktiklerinden oluşur. Tüketicilerin ne istediğini ve ihtiyaçlarının nasıl karşılanacağını bulmak, bu platformlarda süreklilik arz eden bir faaliyettir. Literatür, yeni teorik ve pratik uygulamalarla sürekli olarak geliştirilmektedir. Müşteri satın alma davranışı, tıklama akışı, işlem veya ürün inceleme formları gibi çevrimiçi platformlarda dijital iz verileri bırakır. Bu tez, ürün ve müşteri segmentasyonunu gerçekleştirmek için çevrimiçi işlem verileri üzerinde müşteri davranışı analitiğine yeni bir ağ yaklaşımı sunan bir model önermektedir. Aşağıdaki araştırma sorularına yanıt arıyoruz: Müşteri davranışlarını ve tercihlerini ağ analizi yoluyla anlayabilir miyiz? Birkaç satın alma davranışı türü varsa, bunun altında yatan kalıplar nelerdir? Ağda özel bir rol oynayan belirli özel ürünler var mı? Bu tezin amacı, hedefli reklam, marka sadakatini artırma, arzu edilen müşterileri çekme ve pazarlama mesajlarının etkinliğini artırma gibi pazarlama faaliyetlerini iyileştirme çabalarında karar vericileri destek olmaktır. Gizli ürün topluluklarını keşfetmek için ortak satın alma ağlarında istatistiksel olarak ilkeli bir topluluk tespit yöntemi olan Stokastik Blok Modeli'ni (SBM) kullanıyoruz ve bu topluluklara dayalı iki farklı segmentasyon yöntemi üretiyoruz. Sonuç, geleneksel veri madenciliği yöntemlerini genişleten bir ürün ve müşteri segmentasyonudur. Ürün bazlı segmentasyonu Pazar Sepeti Analizi ile, müşteri segmentasyonunu ise RFM modelleriyle birleştiriyoruz. Modelimizi iki ampirik veri seti üzerinde uyguluyoruz. Son olarak, her iki örnek için de bir yönetici özeti sunuyoruz.

Anahtar Sözcükler: Müşteri Segmentasyonu, Stokastik Blok Modelleme, Ortak Satın Alma Ağları, Topluluk Tespiti, Çeşitlilik

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$P\ (b)$	Prior probability
$P\ (b\|A)$	Posterior probability
$P\ (A\|b)$	Probability of A given b, also called likelihood
$P\ (A)$	Marginal likelihood
θ	Additional model parameters that control how the node partition affects the structure of the network
Σ	Minimum Description Length

LIST OF ACRONYMS AND ABBREVIATIONS

CRM	CUSTOMER RELATIONSHIP MANAGEMENT
EMP	ELECTRONIC MARKETPLACE PLATFORM
MBA	MARKET BASKET ANALYSIS
MCMC	MARKOV CHAIN MONTE CARLO
MDL	MINIMUM DESCRIPTION LENGTH
RFM	RECENCY FREQUENCY MONETARY
SBM	STOCHASTIC BLOCK MODEL
S.D.	STANDARD DEVIATION

1. INTRODUCTION

1.1 Thesis Topic Orientation

Commercial online platforms compete with a multitude of rivals by offering relevant products and services to attract and retain profitable clients. These platforms use a variety of marketing strategies to obtain an advantage over their competitors using their digital trace data (Akter & Wamba 2016). An online marketplace is a platform where multiple third-party companies provide services or commodities. The platforms are essentially responsible for delivering the services that facilitate transactions between their 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. A thorough understanding of consumer purchase behavior is just as critical as analyzing the products and services being provided. Online stores have two main entities: customers and products, each requiring different analysis methods. For instance, while Market Basket Analysis (MBA) is an analysis method for products, the Recency, Frequency, Monetary (RFM) model is widely used to analyze customers.

Market Basket Analysis 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 (Agrawal et al. 1993, Agrawal & Srikant 1994). Although there have been significant contributions from an MBA point of view, there is a limitation on the method's effectiveness (Vindevogel et al. 2005) 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 (Esmaeili & Alireza Hashemi Golpayegani 2021). To address MBA's binary relationship issue, researchers presented a network analysis (Videla-Cavieres & Rios 2014, Kim et al. 2012, Ding et al. 2018) 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 thesis, our research aim is to seek two related outcomes. In the first part, we implement product based segmentation, and in the second part, we carry out a customer segmentation. For the product based segmentation, we empirically analyze 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 bridge or gatekeeper 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 centrality measures to discover products which are significant in terms of business implications: 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 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.

By extending the MBA, this research aims to segment the products by detecting the similarities in customers' co-purchase patterns. The main focus of this research 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. "A network's community structure is uniquely encoded in its wiring diagram" is the fundamental hypothesis (Barabási 2013). SBM is here especially valuable in this context, because it solely works on patterns of this wiring structure.

As for the customer segmentation part of the research, we follow a different path after the community detection phase. Customer segmentation is the practice of effectively grouping a company's customers to target each segment for various marketing activities such as cross-selling, up-selling, and retention in Customer Relations Management (CRM). As markets become increasingly competitive, marketplaces realize that their current customers are their best prospects for existing or new products and services (Kamakura et al. 2003). Cross-selling and up-selling are longstanding and established sales tools (Kamakura 2008) for companies to reach their existing customers. Cross-selling involves selling a similar product related to a previously purchased item, while up-selling involves selling more expensive items.

Analyzing the purchase behavior is a complex and multidimensional process and requires a complex analysis approach for customer retention or increasing customer loyalty. Behavioral analysis of customer segmentation, for example, involves grouping the customers based on traits such as purchase behavior and engagement. The most preferred method for this particular segmentation is the Recency, Frequency, Monetary (RFM) analysis. RFM analysis segments customers based on how much, how frequent, and how recently they purchased the products and services of the organization's platform. It is used for understanding the customers' purchase behaviors; however, despite its usefulness, this approach fails when not combined with other important customer attributes (Tsiptsis & Chorianopoulos 2011).

Although not as common, researchers study customer purchase behavior by analyzing the co-purchase networks (Raeder & Chawla 2009, Ding et al. 2018), which are constructed by connecting two products purchased by the same customer. This research proposes a novel method that combines RFM analysis with Bayesian statistical network analysis based solely on transactional purchase data to perform behavioral customer segmentation. This research intends to improve the segmentation capabilities of the RFM model, especially on the behavioral aspect of its frequency component, by introducing a metric based on principled statistical network methods that detect the diversity of the customers' purchase patterns in co-purchase networks. Our approach determines the diversity by detecting similar purchasing patterns of customers by analyzing co-purchase behaviors with Stochastic Block Model (SBM) community detection methods. We propose to discover the latent product communities by utilizing SBM. This statistically principled community detection method splits the network into product groups based on the similarity of their connection patterns. As the connections in such a network represent the customers' co-purchase, discovering the community structure implies the grouping of the customers based on their purchase patterns.

Frequency is a micro-level metric that only involves a single customer's purchase count independent of other customers' purchases. Diversity is a mesoscale metric that quantifies the number of similar purchase patterns of a customer. While a high frequency indicates the highest number of purchases, the high diversity indicates that the customer has the highest number of different purchase patterns.

Upon examining the community discovery process, our findings show that majority of the customers purchase the products from less than a few communities. In contrast, only a small number of customers purchase from a large number of product communities. This diverse behavior, we believe, should be considered a valuable purchase pattern, indicating a customer who comes to the platform not only for specific products or services. Thus, we define the number of communities of a customer as the diversity score of the customer and combine this metric with the frequency component of the RFM model.

1.2 Thesis Aim

This research aims to segment the products by detecting the similarities in customers' co-purchase patterns. The first focus of this research is to determine the roles of the products in the network and utilize the findings for improving marketing activities such as product recommendation, product placement, cross-selling, or customer retention. Although extensive research has been carried out on segmentation and co-purchase networks, very little research exists which employs statistically principled methods. Despite its many alternatives, SBM is a statistically principled method, making its results domain-independent and less error-prone. Thus, it is scientific technology suitable for decision support systems for any kind of electronic commerce.

The second thesis aim is to propose a customer segmentation model that introduces a novel metric called diversity score. With the intention that the resulting process can contribute to Information Systems by supporting the platform managers in making decisions on marketing campaigns and product promotions. This thesis is aimed to address the following research questions:

- What is the appropriate method that discovers community structure in co-purchase networks?
- What is the most appropriate method that can group products in large networks in a statistically principled approach?
- Is it possible to interpret the inherent structure in the networks as a distinct purchase pattern?
- Product based segmentation
  - Can we distinguish products that play key roles in the network based on their connection pattern and their community characteristic?
  - How can we tackle the issues in traditional co-purchase analysis methods and is it possible to improve these methods by extending them with community detection techniques?

    - How can we interpret network attributes of the discovered communities as business implications?
- Customer segmentation
    - How can we segment customers based on their co-purchase behavior?
    - Can we find customers that the conventional segmentation methods can not notice?
    - How can we tackle the issues in traditional customer segmentation methods and is it possible to improve these methods by extending them with community detection techniques?

1.3 Thesis Outline

This thesis is composed of six chapters. The order of processes followed during the research is as follows:

The second chapter aims to describe the main methodology of this thesis. Figure 2.1 illustrates the proposed framework. This chapter explains the common part of the analytical framework in detail. Chapter three focuses on the product based segmentation part of the thesis. It begins with presenting related works, then follows the theoretical background of the methods employed. The following sections of this chapter provide an empirical implementation of an online platform transaction dataset. Finally, the chapter ends with a discussion of the obtained results.

Chapter four focuses on the customer segmentation part of the thesis. After the related works section, two empirical implementations of customer segmentation are presented. The details of both implementations and results are provided in this chapter for each dataset, followed by respective discussions. This chapter ends with an overall discussion of the customer segmentation process, covering both data set implementations. The published papers during the thesis research are provided in the fifth chapter with their summary and contributions. Finally, the last chapter provides the concluding remarks.

2. RESEARCH APPROACH AND METHODOLOGY

The approach taken in this thesis is a mixed methodology based on network analysis. The main analytical framework is shown in figure 2.1 which consists of three major steps:

1. Data preparation
2. Network construction
3. Segmentation

Two separate segmentation processes follow the network construction step, which results in two different outcomes. In the first step, transaction data obtained from the online platform is cleaned and wrangled to prepare the product and customer entities for the network construction. The second step involves choosing a model for the network, building a bipartite network from the model, and producing the desired projection from the bipartite network. The resulting data structure is a co-purchase network that holds intricate product relationships and customer purchase patterns. Both segmentation steps start with community detection. This thesis frequently uses segment, community, and block terms. Segment is a marketing term, whereas community is a Social Network Analysis term. As for the block term, it is used to refer to the same concept in statistical inference.

One of the essential aspects of the methodology is that the third step yields two different segmentations. The first one leads to product based segmentation, while the second one results in customer segmentation. Additionally, the customer segmentation part introduces a novel metric called diversity. Both processes go through the SBM community detection phase, which is the most crucial element of the segmentation methods.

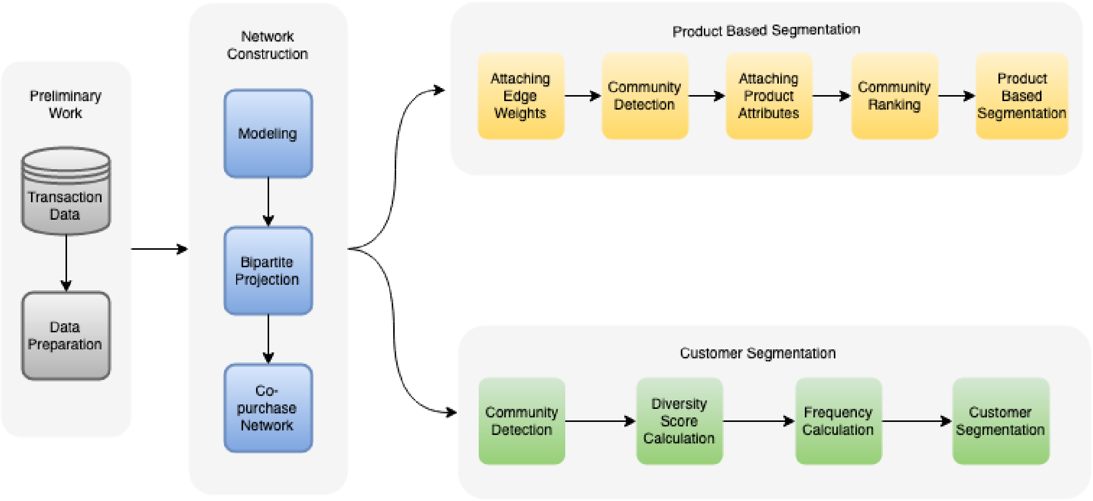

Figure 2.1 Main analytical procedure for product and customer segmentation.

- product based segmentation
  - Before the community detection, the necessary edge weights are attached to the products in the co-purchase network.
  - After determining the parameters, SBM community detection is carried out to discover the latent product groups.
  - The necessary product attributes are calculated and attached to the products.
  - The average attribute values of the products for each community are calculated, and the communities are ranked from top to bottom according to the attribute values.
  - Based on the rankings of their communities, products are divided into segments for marketing activities.
- Customer segmentation
  - Electronic Marketplace Platform Dataset
    - Determining the parameters followed by SBM community detection.
    - Calculating the diversity scores of each customer
    - Calculating the frequency scores of each customer

        * Based on diversity and frequency values, customers are divided into segments for marketing activities.
    - UCI Retail Dataset
        * Determining the parameters followed by SBM community detection.
        * Calculating the diversity scores of each customer
        * Calculating the frequency scores of each customer
        * Based on diversity and frequency values, customers are divided into segments for marketing activities.

In this research, we used the R programming language for the data preparation, manipulation, and network construction processes. For network operations, we used the igraph library (Csardi & Nepusz 2006). Following the construction of the network, community detection, product based segmentation, and customer segmentation processes are carried out in Python programming language with mainly two libraries: a statistical network analysis tool, Graph-tool (Peixoto 2014*a*) and data analysis tool Pandas (McKinney 2010).

2.1 Data Source, Cleaning, and Preparation

This section describes the acquisition and the preparation of the two different data sets used in the research. The raw data set implemented in the product based segmentation belongs to one of the leading Electronic Marketplace Platforms in Turkey and will be mentioned as EMP in this thesis from this point on. It contains nearly 1.5 million transactions, where sellers offer a wide range of products. The transactions took place between 620,767 buyers and 7,516 sellers, involving 412,419 products. The 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 this thesis. Among transactions, a small number of shipping fees that are seen as products had to be removed. In this research, we worked on a portion of the transactions spanning a seven-week time frame, beginning from May 2015, which contains 1,062,925 transactions of 131,951 unique buyers and 178,549 unique products.

The second data set belongs to a UK-based online non-store retail company that mainly sells unique all-occasion gifts. The transactions took place between December 2010 and December 2011. After cleaning the data, 4,175,530 transactions involve 3684 products and 3684 customers. As a point of reference, to keep the comparison of the results consistent, we selected time frames that include a similar number of transactions, which approximates one million for both data sets.

2.2 Modeling the Network

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. Since online marketplace platforms facilitate a transaction between buyers and sellers, the accumulated transaction data contains such entities as buyers, sellers, and products, all suitable candidates for being nodes in a network.

The edges in the network represent the relationship between chosen nodes which can be a transaction 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 focus on this research. Co-purchase here implies that two products are being 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 recommendations. The diagram of the co-purchase network model that is examined throughout this research is shown in Figure 2.2.a. Construction steps of the co-purchase network are explained in section 2.3.

2.3 Extracting the Product-to-Product (Co-purchase) Projection From Bipartite Network

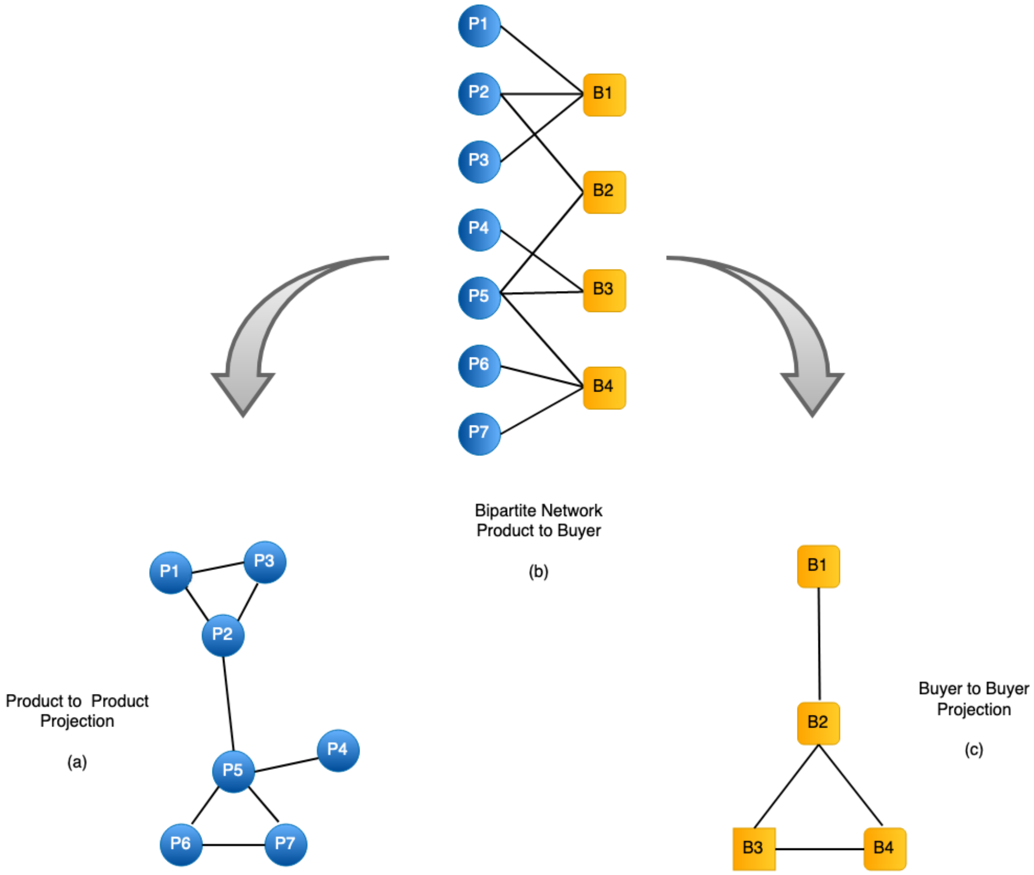

Figure 2.2 Bipartite network model (b). Blue nodes show products and yellow nodes show buyers. Undirected projections of the bipartite network: product-product or the co-purchase network (a), buyer-buyer network(c).

To link two co-purchased products together, 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, and they only connect with the opposing type. Figure 3.b shows a simple model of a product-buyer bipartite network along with two projections on its both sides.

We split the bipartite network into two undirected subnetworks called projections to generate a co-purchase network. One of the projections will be the buyer-to-buyer network, where an edge between two buyers indicates two buyers who bought the same product (Figure 3.c), and 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 2.2.a. Following a similar approach, one can choose other options, such as a product-to-seller bipartite network which can be split into two projections: product-to-product and a seller-to-seller networks. However, we will keep the scope of this research limited to the product-to-product (co-purchase) network illustrated in Figure 2.2.a.

2.4 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 (Holland et al. 1983) when they needed to generate random networks that contain inherent community structure. Later on, scientists ran the algorithm in reverse fashion to infer latent communities within a given network Figure 2.3.

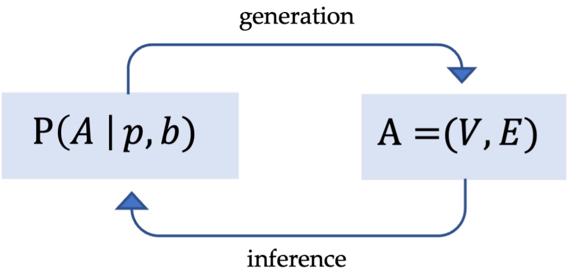

Figure 2.3 SBM generation and inference model. Generation mode: given probability distribution (p) of blocks b, draw network A. Inference mode: given network A (V vertices, E edges) choose p that makes A likely.

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.

We apply a community detection algorithm to this network, which assigns the products to 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 to the rest of the network. In this research, we exploit this similarity concept to segment the products and the customers.

Further inspecting the structure of each community for product based segmentation, 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 their community. Finally, we use the resulting attribute composition to label the community and segment its member products. Chapter 3 provides a detailed explanation of each step of the product based segmentation process.

Examining a customer's co-purchase, the two products may either fall into the same community or two different communities. If the co-purchase of a customer connects two separate communities, we see this as a diverse purchase behavior. Then, we call the total number of a customer's diverse co-purchases as their diversity score. Based on this novel metric and the purchase frequency, we perform customer segmentation. Chapter 4 provides a detailed explanation of each step of the product based segmentation process.

2.4.1 Generative Mode 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 $b_i \in \{1, ..., B\}$ entries that represent the building blocks of the network. Given the above information SBM generates a network with equation 2.1 and 2.2, where μ is the lagrangian multiplier and P_{rs} is the probability of existence of an edge between two nodes from groups r and s.

$$P\left(A|\ p, b\right) = \prod_{i<j} P_{b_i,b_j}^{A_{ij}} \left(1 - P_{b_i,b_j}\right)^{1-A_{ij}} \tag{2.1}$$

$$P_{rs} = \frac{e^{-\mu_{rs}}}{1 + e^{-\mu_{rs}}} \tag{2.2}$$

2.4.2 Inference Mode of SBM, Bayesian Inference

The inference mode of the SBM is the Bayesian Inference. Bayesian statistics is a method for analyzing data that is based on Bayes' theorem. In this method, existing knowledge about the parameters in a statistical model is updated with the information in observed data. In the field of statistics, the Bayesian inference method is an essential tool. Bayesian updating is an especially important technique to utilize. Bayesian inference has been utilized in a wide variety of fields, including the social sciences, genetics, medicine, engineering, and ecology.

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 2.3 where $P\ (b|A)$ is the posterior distribution. This modeling approach makes this method a principled method rather than a heuristic one.

$$P\ (b|A) = \frac{P\ (A|b)P\ (b)}{P\ (A)} \tag{2.3}$$

$P\ (b)$ is the prior probability.
$P\ (b|A)$ is the posterior probability.
$P\ (A|b)$ is the probability of A given b, also called likelihood.
$P\ (A)$ is the marginal likelihood.

We can simply explain the inference mode of the SBM as, "When we observe a network, what is the likelyhood that it was generated by the given community structure via the Bayesian posterior probability."

Equation 2.3 can be written as equation 2.4 then, Σ gives the Minimum Description Length (MDL) which is used to determine the iterarion parameters of the algorithm. It determines how much information is necessary to explain the data if we encode it using a specific parametrization of the generative model. Here, "The simplest model is selected, among all possibilities with the same explanatory power. The selection is based on the statistical evidence available, and therefore will not overfit." (Peixoto 2019).

$$P\ (b|A) = \frac{Exp(-\Sigma)}{P\ (A)} \tag{2.4}$$

$$\Sigma = -\ln P\ (A|\theta, b) - \ln P\ (\theta, b) \tag{2.5}$$

where θ are additional model parameters that control how the node partition affects the structure of the network.

“When analyzing empirical networks, one should be open to the possibility that there will be more than one fit of the SBM with similar posterior probabilities. In such situations, one should instead sample partitions from the posterior distribution.” (Peixoto 2020). For this reason, we utilize the Markov Chain Monte Carlo (MCMC) algorithm. In this method, nodes are moved into different groups with varying probabilities, and these moves are either accepted or rejected so that, over time, the desired partition probabilities can be observed. Due to the fact that each MCMC sweep is independent of the number of groups used in the model and has a run-time that is linear with network edge count, the algorithm is applicable to large networks.

2.4.3 SBM Types

Although there are several versions of SBM, we employ a combination of three of its versions in this research: degree corrected SBM (Karrer & Newman 2011), Hierarchical SBM (Peixoto 2014*b*), and weighted SBM (Aicher et al. 2015, Peixoto 2018). The standard SBM assumes that the probability of nodes connecting to each other within a community is equal, which does not agree with most of real-world networks. This assumption makes the standard method sensitive to high degree nodes. Karrer & Newman (2011) 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, the 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 2.6.

$$P\ (b|A,x) = \frac{P\ (x|A,b)\ P\ (A|b)\ P\ (b)}{P\ (A,x)} \tag{2.6}$$

Where x is a model for weights between blocks, the algorithm allows us to choose weight models from exponential, normal, and binomial options depending on the type of data.

2.5 Segmentation Methods Derived From Discovered Communities

After discovering the communities, the research method follows two separate paths involving a product and a customer segmentation. Figure 2.4 illustrates the product based segmentation based on nodes of a community and the customer segmentation based on the pattern of the edges between communities. Because it is not possible to see details in the visualization of a large network of millions of edges, as a representative model, we have illustrated the co-purchase network of one-day transactions in Figure 2.4. The following two chapters describe each segmentation process's theory, implementation, results, and discussion.

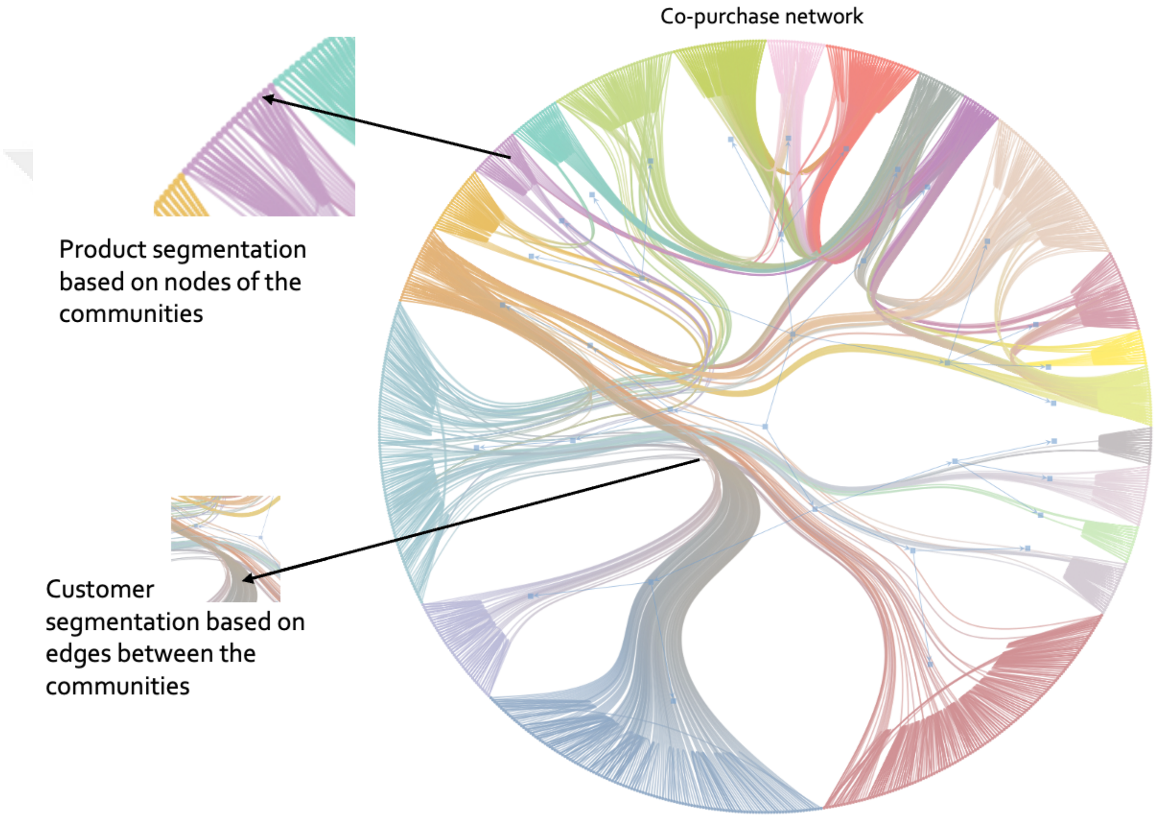
Co-purchase network
Product segmentation based on nodes of the communities
Customer segmentation based on edges between the communities

Figure 2.4 The map of a representative co-purchase network. Colors indicate different communities detected by the SBM. Two groups of the research area after community discovery.

3. PRODUCT BASED SEGMENTATION

Figure 3.1 illustrates the proposed framework of this research for customer segmentation. It consists of data preparation, network construction and community detection, diversity score calculation, and customer segmentation. We have employed the processes in this framework in two empirical data sets: a leading marketplace platform in Turkey, EMP, and UCI retail data set from a UK-based online store (Chen et al. 2012). We used the number of transactions as a frame of reference to compare the research findings.

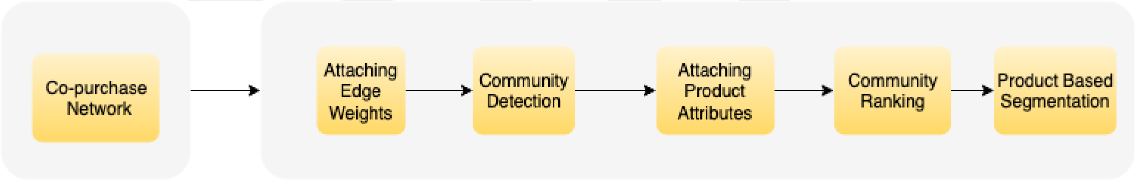

Figure 3.1 Subsection of the analytical procedure for the product based segmentation.

3.1 Related Works

Market Basket Analysis (MBA) is considered the most common way to understand co-purchase behavior both in the industry and in academia (Büchter & Wirth 1998, Woo 2013). Agrawal et al. (1993) 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. (2013) incorporate k-means clustering algorithm into the MBA to perform product based 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

& Jedrusik (2021) 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. (2018) point out the lack of network understanding.

(Raeder & Chawla 2009) describe the issue as:
“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 since traditional Market Basket Analysis is often difficult to isolate interesting relationships.”

Ding et al. (2018) 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 (Kim et al. 2012). Many researchers go further and add community detection to the research (Raeder & Chawla 2009), 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” (Peel et al. 2017, McCarthy et al. 2019), 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 (Ghasemian et al. 2019) has a resolution limit that prevents it from detecting small communities in large networks (Fortunato & Barthelemy 2007).

Table 3.1 A summary of literature in terms of three criteria involving customer segmentation with community detection or clustering.

Researchers	**Research Focus**	**Analysis Method**	**Attribute used for Segmentation**	**Community Detection Methods or 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

Researchers	Research Focus	Analysis Method	Attribute used for Segmentation	Community Detection Methods or Heuristics used
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 based 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 (Kim et al. 2012, Ma'arif & Mulyanto 2014, Oestreicher-Singer et al. 2013, Faridizadeh et al. 2018). However, much of the research has either applied problematic detection methods such as modularity maximization (Newman 2006) or focus on basic centrality measures or clustering behaviors to analyze the network (Kim et al. 2012, Faridizadeh et al. 2018, Huang et al. 2007). The study of Raeder and Chawla Raeder & Chawla (2009) is one of the early examples of using a 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. (2012) take a similar dataset of transaction data from a department store and model two different co-purchase networks. One connects two products if they appear on 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 & Rios (2014) 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 (Raeder & Chawla 2009) their method involves filtering edges to reduce the network to manageable sizes; however, removing edges of a network might compromise the underlying network structure. In this thesis, we cover the entire transaction data. Moreover, contrary to many studies (Videla-Cavieres & Rios 2014, Kim et al. 2012) 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 (Peixoto 2019) 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. (2018) employ SBM among the studies seen in Table 3.1. 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 (Peixoto 2014*b*). 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. (2019), and Chattopadhyay et al. (2020) 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 afterward.

There are various methods to achieve product based segmentation. Artificial neural networks are a recent example. Wang et al. (2019) use Self Organizing Map an artificial neural network method to segment the products. Additionally, they incorporate Recency, Frequency, Monetary analysis into their research. Apart from co-purchase analysis, product based segmentation can be performed based on demographic data. For instance, Lees et al. (2016) present demographic product based 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, In this thesis, we perform a behavioral product based segmentation.

3.2 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 research, we group similar products and 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 (Freeman 1977) 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 eigenvector 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 (Newman 2008) captures this nuance quantifying the centrality of a vertex accordingly.

3.3 Attaching the Edge Weights

The product-product (co-purchase) network is undirected, meaning 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 the 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 3.2 is a simple model showing how the attributes are attached to the network on both nodes and edges. A list of various possible information that can be used as a node or edge attribute extracted from the transaction data is shown in Table 3.2. However, in this thesis, 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 3.2 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 research can be carried out using the frequency of the purchases as the edge weights that can reflect differently on the research findings.

Table 3.2 List of likely edge weights and node attributes that can be extracted from the transaction data.

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

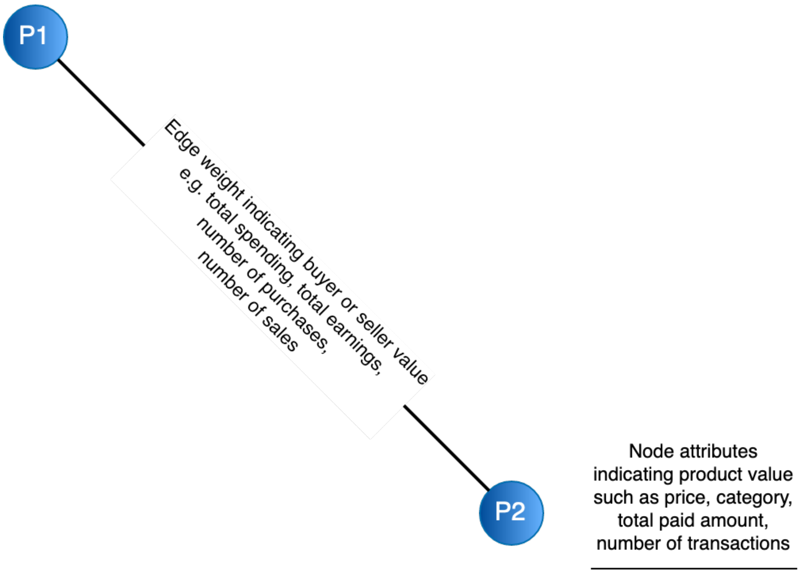

Figure 3.2 Co-purchase network model showing potential edge weights and node attributes. Only the amount of money spent is used in the research (monetary attribute).

3.4 Parameter Selection for Community Detection

For finding a good estimate for community detection we run a greedy algorithm based on merge-split Markov chain Monte Carlo (MCMC) (Peixoto 2020). We performed several runs with a varying number of Monte Carlo sweeps and iterations beginning from one week up to seven weeks of co-purchase networks. We then plotted the Minimum Description Length 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. Appendix A provides a selection of plot of these trials Figures A.1, A.2, A.3.

3.5 Attaching Product Attributes

Up to now, the nodes have no attributes other than their product ids. We calculate the betweenness and eigenvector centrality scores of each product in the network. Additionally, the monetary attribute of each product is attached to the network. Naturally, the attributes exhibit varying ranges of values; for instance, the betweenness score always ranges between 0 and 1, whereas the monetary attribute may have a wide total price range. To be able to compare their values, we calculate the rank of each value using the 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.

3.6 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 since 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.

3.7 Implementation and Results

The SBM algorithm discovered 309 product communities, and computation time took one hour 32 minutes to complete with 10 MCMC sweeps per iteration and 200

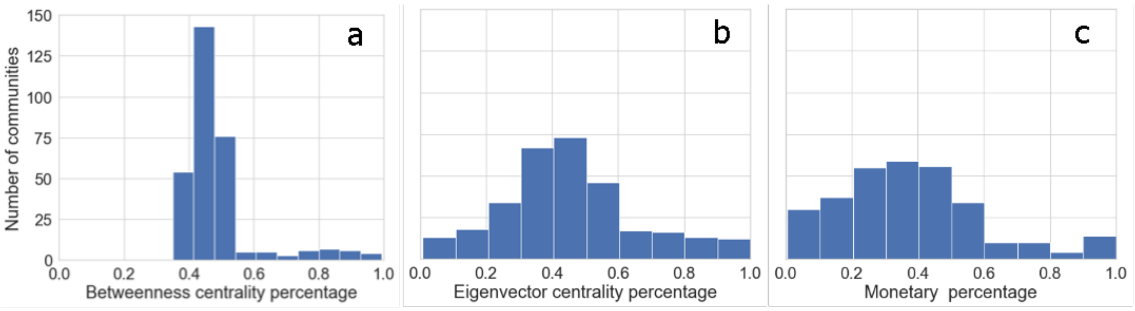

Figure 3.3 Mean percentage histogram of attributes vs number of communities.

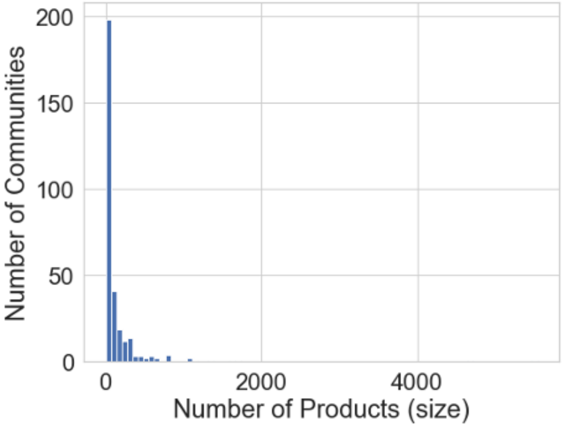

Figure 3.4 Mean percentage histogram of attributes vs number of communities.

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 3.3 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 3.3.

Examining the betweenness attribute in Figure 3.3 a, we observe that none of the communities have a mean percentage lower than 35%, and many communities lie between 35% - 55%. The rest have very low values, and they are almost equally distributed. The Eigenvector centrality is close to a normal distribution Figure 3.3.b. As for the monetary attribute Figure 3.3.c., the range between 30% and 50% has the largest number of communities.

There is a community with 5,543 products, another with 4,683, and the following largest six communities contain between 1,000 and 2,000 products. We use the community size histogram (Figure 3.4) to determine the cutoff thresholds for level labels; small, medium, and large. Table 3.3 is a list of the cutoff points determined by examining their distributions.

Table 3.3 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

Table 3.4 shows the breakdown of the number of community attributes which is determined by the thresholds given in Table 3.3. 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 3.4 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 3.5 is the correlation matrix of the community attributes. The betweenness attribute highly correlates with the monetary attribute.

Figure 3.5 shows four representative communities with various sizes and characteristics. The details of the communities in Figure 3.5 are shown in Table 3.6, listing

Table 3.5 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

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 the 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 3.5.a has high levels in all attributes, and there are 14 similar communities of various sizes. The community in Figure 3.5.b exhibits similar values with one difference that the standard deviations are much smaller. One of the largest communities in the network (Figure 3.5.c) 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 of a dominant attributes is the community in Fig 3.5.d 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 section of the co-purchase network is shown in Figure 3.6 where there are two main product groups, groceries and mobile phone accessories. Milk and phone case having high eigenvector centrality values, whereas phone charger having high betweenness centrality value.

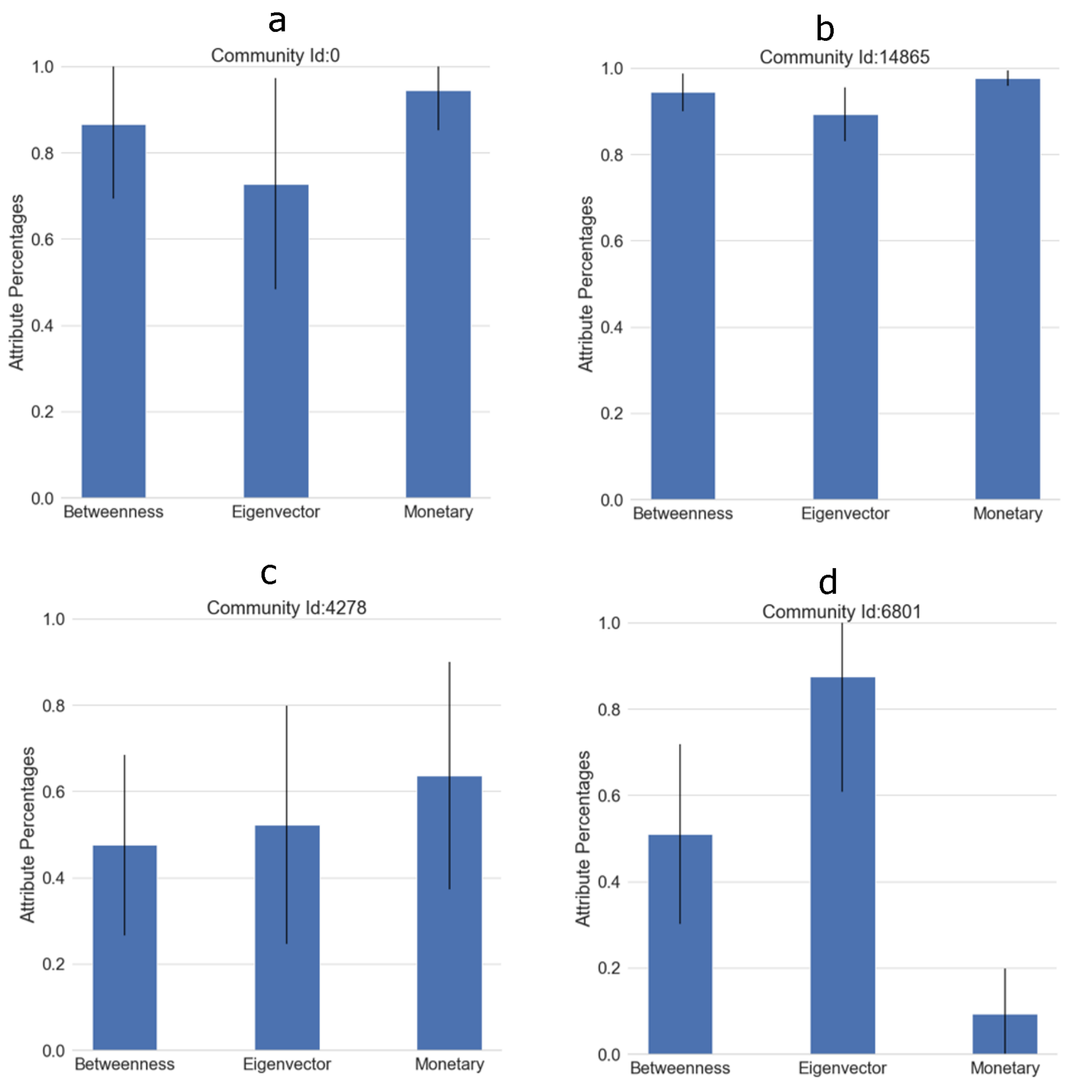

Figure 3.5 Mean attribute percentages and standard deviations of four selected communities.

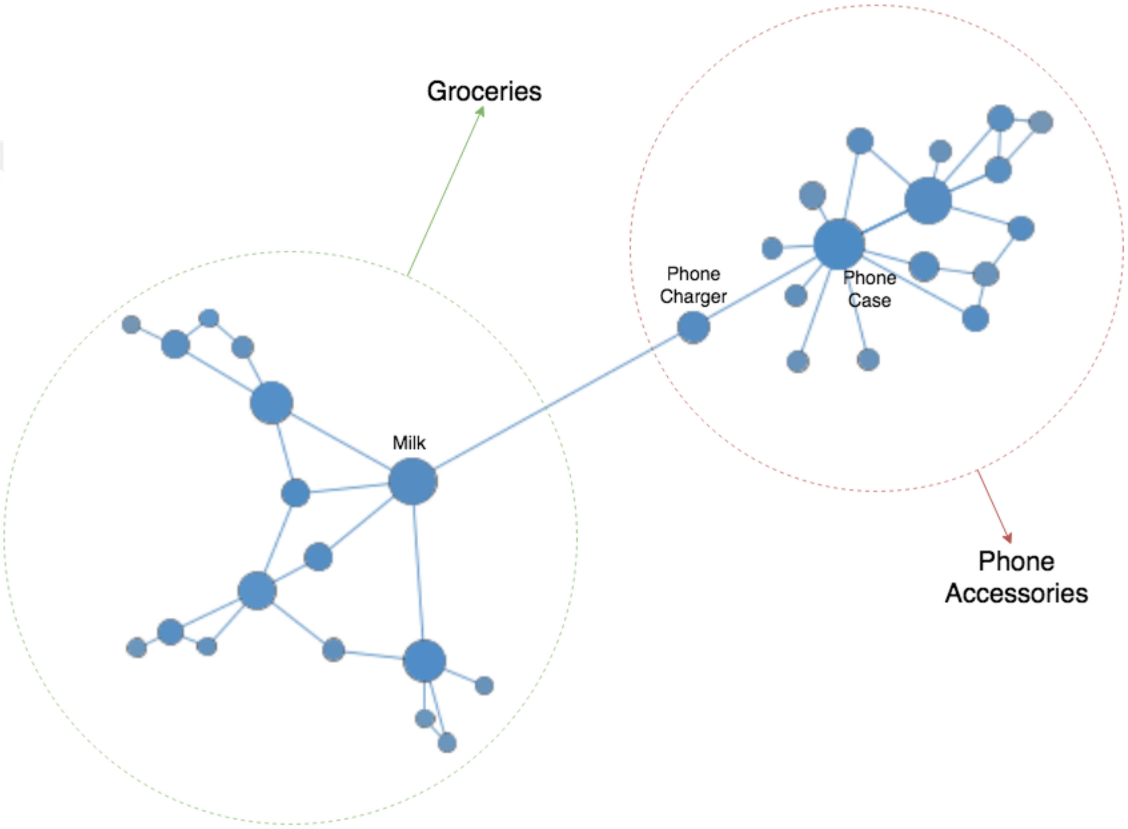

Figure 3.6 A subgraph of the co-purchase network showing two product groups with high centrality products (milk and phone charger).

Table 3.6 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

3.8 Discussion

In this section of the thesis, 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.

The Modularity Maximization community detection method can find communities in a network even if there are no underlying communities in the network (Guimera et al. n.d.). One of the prominent features of the SBM is that it can detect whether the network has a community structure or not. The SBM's Bayesian inference is built from the ground up to avoid the issue in a principled manner, and it consistently succeeds (Peixoto 2019).

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 since we used this monetary attribute as the edge weight of the SBM algorithm. The correlation matrix in Table 3.5 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 since 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 3.4. 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 1,175 products which exhibits high levels in all attributes (Fig 3.5.a). 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 3.5.b).

A monetary dominant community (Figure 3.5.c) 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. (2018) 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 thesis, we find the communities that contain products with high eigenvector centrality values. The community in (Figure 3.5.d) 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 them 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. (2018) 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.” (Ding et al. 2018). In terms of business implications, their study concludes that segmenting products by their role in the network will help marketers to develop effective strategies for 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 customer who purchases groceries can be recommended a phone charger. 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.

4. CUSTOMER SEGMENTATION

Figure 4.1 illustrates the proposed analytical framework subsection of this research for customer segmentation. It consists of data preparation, network construction and community detection, diversity score calculation, and customer segmentation. We have employed the processes in this framework in two empirical data sets: a leading marketplace platform in Turkey, EMP, and UCI retail data set from a UK-based online store (Chen et al. 2012). We used the number of transactions as a frame of reference to compare the research findings.

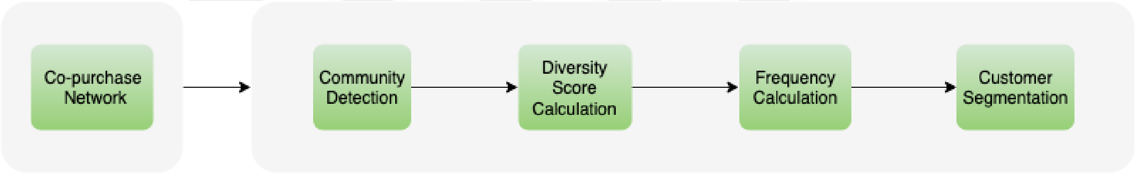

Figure 4.1 Subsection of the analytical procedure for the customer segmentation.

4.1 Related Works

A large and growing body of literature has investigated customer segmentation. Although researchers and business professionals arrange customer segmentation into a varying number of categories, mainly four types are widely used in business and academia: geographic, demographic, psychographic, and behavioral. Behavioral segmentation can be based on spending habits, customer loyalty, or customers' action on the company website or application. This research focuses on detecting similar purchasing patterns of customers by analyzing co-purchase networks with community detection methods. There are a handful of studies applying such an approach to customer segmentation. Table 4.1 shows a comparison between seven representative studies.

Helal et al. (2016) focus on customer segmentation for viral marketing activities. They propose an analysis method that identifies the most influential actors in the network and uses them as seeds to expand their communities. Wang et al. (2017) also employ a network community detection approach to customer segmentation. Customers' product ratings present ample information on customers' opinions about the products. Based on that assumption, they propose a novel method that detects communities. The method splits the network into "sentiment communities." The study uses the customer segments for target marketing.

Previous studies examine the social media interactions with network community detection algorithms for customer segmentation, link reference, and product recommendations (Suryateja & Palani 2017). For instance, Wang et al. (2017) incorporate the customers' social media interactions into customer segmentation. The interactions occur in the company's social media platforms used to build an interaction network. The study focuses on faster market segmentation, and to achieve this goal, it discovers communities in the network using a modularity maximization method. Ballestar et al. (2018) also incorporate customer interactions in the social network of the company's website. They focus on increasing the long-term profitability and loyalty of the customers by showing how the customer's role within the social network determines the customer's commercial behavior.

The segmentation results sometimes might be challenging to interpret for the managers. To address this issue, Korczak et al. (2019) employ a multi-level method involving a label propagation algorithm combining RFM analysis with the K-NN clustering method. Shi-Yong et al. (2019) study how the community structure plays a role in the diffusion of knowledge and compare model selection for community detection and seed selection strategy. Customer segmentation can be challenging on massive data sets. Zhang et al. (2021) propose utilizing a bipartite modularity maximization algorithm to address this issue. They compare the results with the

RFM model. This research focuses on customer segmentation based on the diversity of their purchase patterns compared to the frequency component of the RFM model by using the nature of the SBM community detection algorithm, which allows us to identify stochastically similar purchase patterns in purchase transaction data.

Table 4.1 A summary of literature in terms of three criteria involving customer segmentation with community detection or clustering.

Researchers	**Research Focus**	**Analysis Method**	**Community Detection or Clustering Method**
Helal et al., 2016	Customer segmentation for viral marketing	Community detection based on finding the most influential actors of a network	A novel method based on influence propagation
Wang et al., 2017	Customer segmentation, target marketing	Analysis based on customers' product ratings	A novel method called sentiment community detection
Alamsyah, 2017	Faster market segmentation	Detecting communities from customers' social media interactions	Modularity Maximization
Ballestar et al., 2018	Increase customer loyalty and long-term profitability.	Clustering the customers based on their commercial and social activities	Agglomerative hierarchical clustering based on log-likelihood distance
Korczak et al., 2019	Increase the definition of the customer communities	Combine community detection with RFM and K-NN	Density-based clustering with a label propagation algorithm

Researchers	**Research Focus**	**Analysis Method**	**Community Detection or Clustering Method**
Shi-Yong et al., 2019	Market segmentation based on diffusion of knowledge	Analyzing the efficiency of knowledge diffusion in communities	Modularity Maximization
Zhang et al., 2021	Customer segmentation for cross-selling reduces computational complexity	Community detection in bipartite graph and compare results with RFM model	Modularity Maximization
This Research	Customer segmentation based on diversity score, improving Frequency metric in RFM model	Community detection and customer diversity	Degree-corrected Hierarchical Weighted SBM

4.2 Theoretical Background

4.2.1 Customer Segmentation

Finding the target customers is a crucial part of a marketing campaign process. One cannot aim at the company's entire customer base and direct all types of marketing messages to them. The messaging should be optimized for the appropriate audience for the desired marketing activity. Otherwise, many risks emerge, such as wasting limited resources or exhausting the attention of valuable customers. Customer segmentation is a data mining technique in CRM that helps determine the target customers for the intended marketing campaigns by grouping the customers based on specific characteristics. Table 4.2 shows four main categories of customer segmentation and the related customer characteristics for each category.

Table 4.2 Four main types of customer segmentation.

Geographic	**Demographic**	**Psychographic**	**Behavioral**
Country	Age	Lifestyle	Benefits Sought
City	Gender	AIO. Activity	Purchase
Density	Income	Interest, Opinion	Usage
Language	Education	Concerns	Intent
Climate	Social Status	Personality	Occasion
Area	Family	Values	Buyer Stage
Population	Life Stage	Attitudes	User Status
	Occupation		Life Cycle Stage
			Engagement

4.2.2 Recency Frequency Monetary (RFM) Analysis

The extent of this research falls under the behavioral customer segmentation category, mainly focusing on the purchase and engagement subcategories. The most preferred method is the RFM analysis to understand the customer's purchase behavior (Tsiptsis & Chorianopoulos 2011). This model consists of three metrics: recency (the most recent purchase of the customer), frequency (total number of customer transactions), and monetary (total or average value of the transactions). These metrics quantify the key traits of the customers. For instance, the more recent purchase indicates more responsiveness to promotions. Higher purchase frequency suggests more engagement and satisfaction. The monetary factor is a good indication of the customer's purchasing power. The origin of this technique goes back to the catalog industry in the 1980s, where they successfully used it to target the right customers in marketing campaigns. "Although useful, the RFM approach, when not combined with other important customer attributes such as product preferences, fails to provide a complete understanding of customer behavior" (Tsiptsis & Chorianopoulos 2011).

4.3 Community Detection Algorithm paramaters

A greedy technique based on merge-split Markov chain Monte Carlo (MCMC) is used to get a good approximation for community discovery (Peixoto 2020). For both data sets with various time intervals, we ran numerous runs with variable numbers of Monte Carlo sweeps and iterations. We then displayed the entropy for each iteration to follow the minimization process and determined the optimal number of iterations for chain equilibration. For both data sets, we opted to run the method with ten sweeps for 200 iterations.

4.4 Calculating the Customer Diversity Scores

The community discovery process (SBM) assigns each product to a community based on the customers' co-purchase patterns. Each co-purchase of a customer is represented by a connection between two products in the network. Most of these connections occur between products of the same community, yet few occur between products belonging to different communities. We identify the diversity score of a customer as "one" if the products that the customer purchased fall into the same community. This lowest value indicates that the customer's purchase pattern is monotonous. Following this logic, we calculate a customer's diversity score by summing up the number of product communities. A higher diversity score, therefore, indicates a more heterogenous co-purchase pattern. We plot the log-log distribution of the diversity score to compare them for each network.

4.5 Comparison of Frequent and Diverse Customers

From the three parameters of RFM analysis, frequency has the closest relation to the co-purchase concept. Presumably, the diversity score ought to correlate with purchase frequency because a customer has to make a high number of purchases to have a high diversity score. However, the diversity is not the same as the frequency. To quantify the relationship between them, we look at the correlation coefficient between the frequency (number of purchases of a customer) and the diversity score. In other words, to distinguish the purchase frequency from the community diversity, we examine their correlation coefficient. To further the inspection, we draw the scatter plots of the two variables.

4.6 Customer Segmentation Based on Purchase Frequency and Diversity Scores

Customer segmentation is implemented by distinguishing the contrast between the diversity and frequency values of the customers. To achieve this, we plot the purchase frequency (y-axis) against the customers' diversity score (x-axis) and split the plot into quadrants. Based on these quadrants, we segment the customers ending up with four labels: high diversity high frequency, high diversity low frequency, low diversity high frequency, and low diversity low frequency. Marketing managers can arbitrarily determine the levels as high, medium, and low based on their domain requirements according to the Pareto principle.

4.7 EMP Dataset Implementation and Results

Figure 4.2 shows diversity scores and their corresponding number of customers on a logarithmic scale. SBM splits the network into communities in such a way that 186,443 customers purchased products from only one community, while only one customer purchased products from 14 different communities (Table 4.3). Presuming low diversity scored customers as monotonous customers and the others as diverse, we see that majority of the customers' products fall into only one or two communities. The SBM algorithm is supposed to achieve this since the products are divided into communities based on their customers' co-purchase patterns. However, we are interested in high diversity customers who purchased products from several communities.

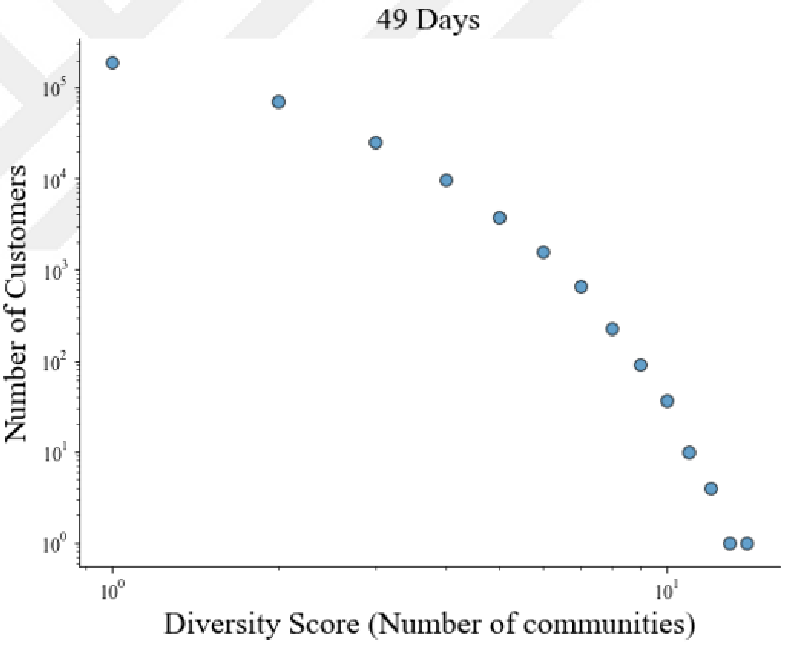

Figure 4.2 The distribution of diversity scores of the customers on a logarithmic scale.

Table 4.3 Number of customers for each diversity score.

Diversity Score	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Number of Customers	186,443	70,732	25,087	9,655	3,812	1,577	663	227	89	36	10	4	1	1

Figure 4.3 Graph shows how co-purchases of the top three customers connect several

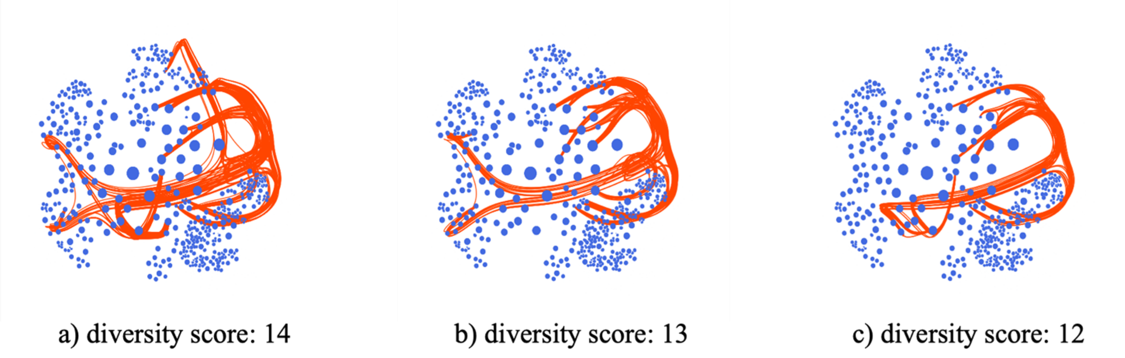

Figure 4.3 Three graphs showing how co-purchases of the top three customers connect several communities. The blue dots show product communities, and their sizes indicate the number of products. Red lines show the co-purchases of the customer.

communities. The blue dots show product communities, and their sizes indicate the number of products. Red lines show the co-purchases of the customer.

The inter-community connection map of the top three diverse customers seen in Table 4.3 is illustrated in Figure 4.3. We can call an arbitrary number of top customers as champions in terms of diverse purchasing. The highest score belongs to a customer who connects 14 different communities (panel a). The second highest diversity score belongs to a customer who connects 13 communities (panel b), and four customers share the third place in diversity scores. Panel (a) shows one of them representing the customers with a diversity score of 12. Notice how these customers have a similar pattern showing common communities, which turns out to be mostly supermarket, technology-related products, or another group of similar patterns with a different set of common product communities. The size of the blue dots indicates the number of products in that community. A quickly noticeable observation is that all top customers connect relatively large communities, although a few of them are smaller in size.

Table 4.4 shows correlation coefficients between the diversity score and the RFM metrics. The recency and monetary metrics show a relatively low-level correlation with the diversity metric. The reason recency has a negative correlation is that it

Table 4.4 Correlation coefficients of RFM metrics and diversity score.

	Recency	**Frequency**	**Monetary**
Diversity Score	-0.230	0.710	0.282

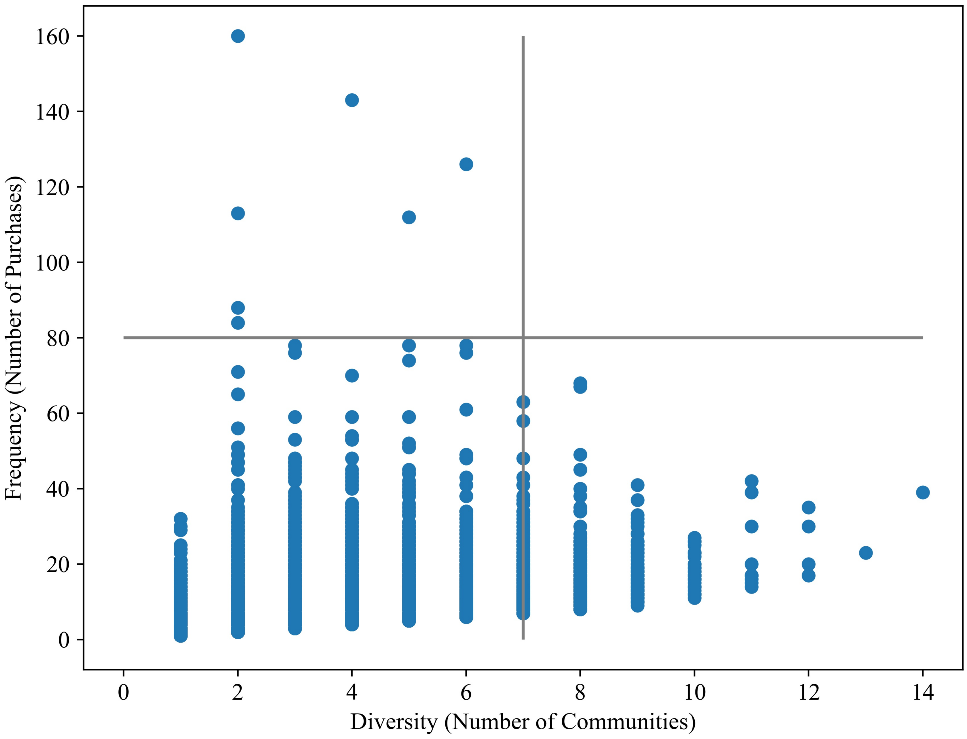

Figure 4.4 Diversity score vs. purchase frequency in EMP data set.

shows the number of days passed since the customer's last purchase. Therefore, a customer having a recent purchase has a low recency value. As for the frequency metric, it shows a much higher correlation of 0.710, which is why we focus on this metric and ignore the other two metrics for performing customer segmentation.

We examine the frequency, which is the number of purchases of a customer, and compare it with the customer's diversity score. Figure 4.4 shows the scatter plot of the two variables. The customers with the highest diversity scores are not frequent customers, and the ones with the highest purchase frequency are low diversity score customers. The majority of the customers are in low frequency and low diversity zone. There are no customers who purchase both frequently and diversely.

4.8 EMP Discussion

SBM finds that this co-purchase network has a community structure divided into 510 separate product communities. The highest diversity score belongs to a customer who made purchases from 14 of these 510 communities. We have ignored recency and monetary components and focused on the frequency based on the premise that diversity is closely related to the frequency and not correlated to that extent with recency and monetary. Correlation coefficients in Table 4.4 support that claim as the correlation with frequency are greater than 0.7, and the others are less than 0.3.

Our method segments the customers based on the contrast between their purchase frequency and purchase diversity behaviors. To achieve that, we separate the customers into four segments, shown as the quadrants in Figure 4.5, illustrating a selection of customers in red. There are four different types of customers in terms of the two metrics. The majority of the customers fall in the low frequency, low diversity segment. On the other hand, there are no customers in this network's high frequency, high diversity segment. The RFM model easily detects customers in these two quadrants, where the marketers target the former and completely ignore the latter.

There are seven customers in the high frequency, low diversity segment. Four of those customers shown in red are frequent buyers, yet their diversity levels are almost at the bottom. Frequency analysis would put them on the target list; however, marketers should ignore them due to their low diversity scores. The last segment belongs to low frequency, high diversity customers, which would not be noticed with the frequency analysis. This segment contains customers under the radar, yet their top diversity scores indicate that they connect different types of products from several communities, increasing the platform's integrity. Table 4.5 explains the

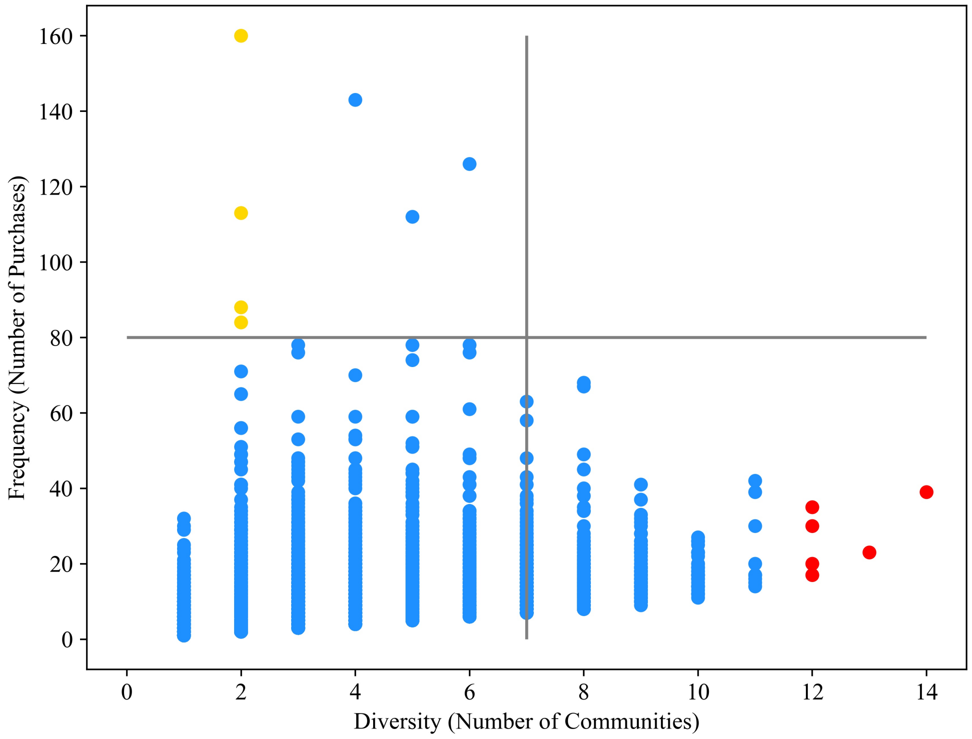

Figure 4.5 Diversity vs. Frequency of the EMP network.

two significant segments shown in Figure 4.5, the second and fourth quadrants. The first quadrant, which contains no customers, and the third quadrant, which contains customers with low values for both attributes, are not included in the table.

Table 4.5 Business implications of the two significant segments of EMP customers.

Customer Segment	Second Quadrant (High frequency - Low diversity, top-left)	Fourth quadrant (Low frequency-High diversity, bottom-right)
Business implications	Four of seven customers are frequenters, but not diverse customers (shown as yellow in Figure 4.5) can be ignored. In contrast, three of seven (blue) are future candidates to become high diversity customers. Frequent customers are generally overwhelmed by inefficient marketing campaigns. This segment helps sharpen the focus of the marketing managers.	The six of the customers shown in red in this segment are under the radar of the RFM model due to low-frequency values. On the other hand, they are the top diverse customers on the platform; thus, they are good candidates that give us an additional shortlist to target when issuing campaigns to increase loyalty.

4.9 UCI Retail Dataset Implementation and Results

We examined the UCI retail transactions in each quarter of the year separately. Figure 4.6 shows the customers' diversity distribution in the year's first quarter. The rest of the quarters exhibit similar distribution characteristics. Like the previous data set, the high diversity customers are rare, and the majority are low diversity customers.

Table 4.6 shows the number of communities SBM inferred for that period, with a maximum value of 176 and a minimum value of 137 in the third and fourth quarters, respectively. We examined the correlation between the RFM metrics and diversity values. We see similar results as the previous data set. Once again, the recency

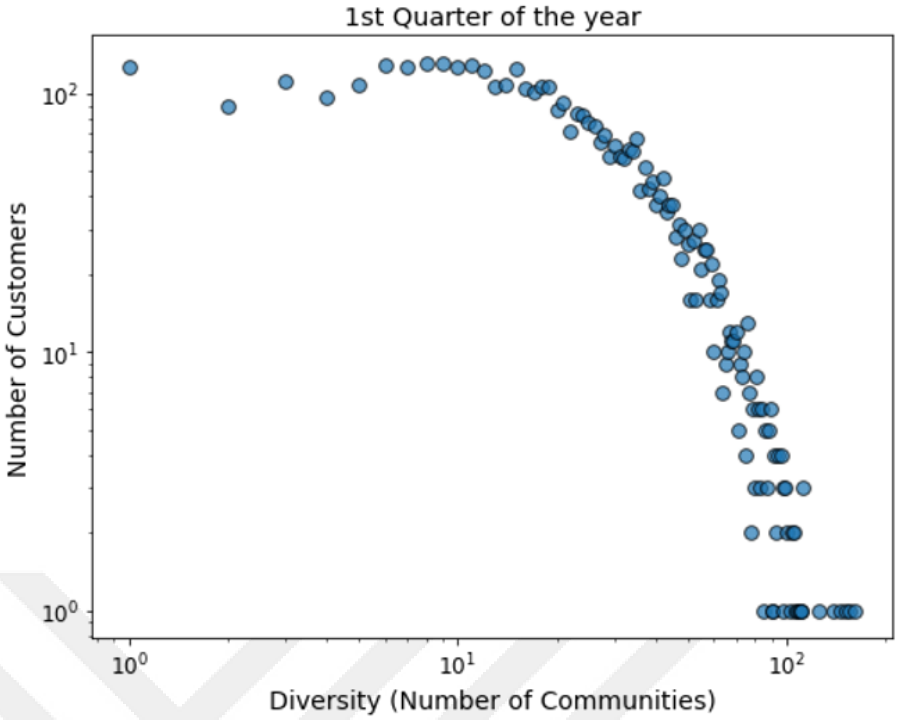

Figure 4.6 The distribution of diversity scores of the customers on a logarithmic scale.

has small negative values for all the quarters and small monetary positive values. In contrast, the frequency shows much higher values between 0.586 and 0.627 compared to the other metrics.

Table 4.6 Correlation coefficients of customer diversity vs. purchase frequency and basic attributes of the network for each quarter of the year.

Quarters of the year	**Q1**	**Q2**	**Q3**	**Q3**
Corr. Coeff. Diversity vs. Recency	-0.348	-0.362	-0.384	-0.383
Corr. Coeff. Diversity vs. Frequency	0.627	0.610	0.606	0.586
Corr. Coeff. Diversity vs. Monetary	0.280	0.283	0.274	0.282
Number of products	2,274	2,761	2,820	3,012
Number of transactions	1,081,199	1,128,052	1,216,159	2,189,702
Number of communities	175	173	176	137

Figure 4.7 shows how frequently the customers made purchases from the platform

(frequency) versus how many different communities the purchased products belong to (diversity). Each figure separates the customers into four quadrants. All of the periods exhibit similar characteristics. For instance, low frequency, high diversity zone has no customer in all periods. Whereas most of the customers are in the low frequency, low diversity zone. There are only a few customers who are frequent and diverse customers at the same time. The rest of the customers reside in the low-frequency zone, yet their diversity is high. Table 4.6 shows the number of communities SBM discovered for that period, with a maximum value of 176 and a minimum of 137 in the third and fourth quarters, respectively.

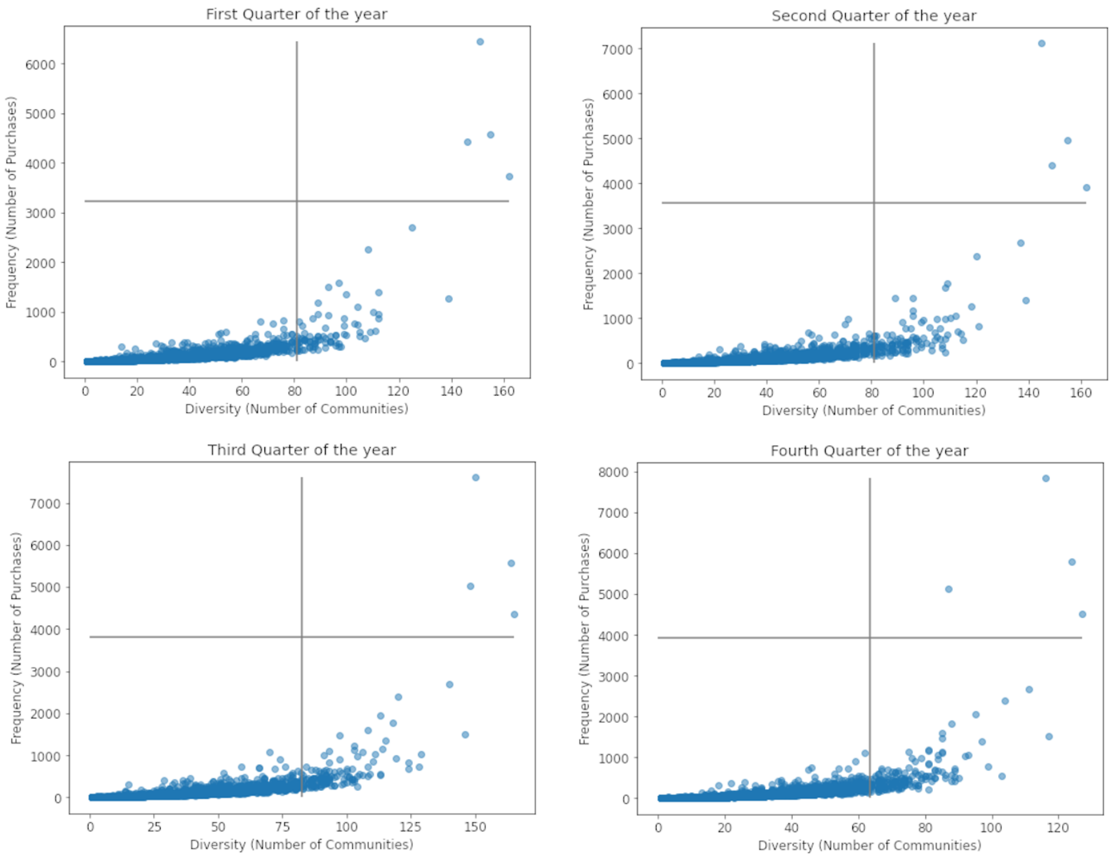

Figure 4.7 Diversity vs. frequency in UCI retail dataset.

4.10 UCI Retail Discussion

SBM can distinguish community from random structure; in this case, it infers that the product network has a community structure. In the third quarter of the year, the co-purchase network contains 176 separate product communities. The highest diversity score in this network belongs to a customer who made purchases from 165 of these communities. Based on the frequency and diversity of their purchases, we draw a plot and separate the customers into four segments by dividing the plot into four quadrants. In other words, we segment the customers into four main categories based on the contrast between these diversity and frequency metrics. Figure 4.7 shows the quadrants which are named counter clock-wise starting from top-right.

Frequency is a part of RFM analysis commonly employed in academia and industry. Naturally, a customer has to make a high number of purchases to make purchases from different communities. This makes diversity correlate with frequency. However, the frequency is not entirely correlated with the diversity. For this dataset, it is a moderate value of 0.6 (Table 4.6).

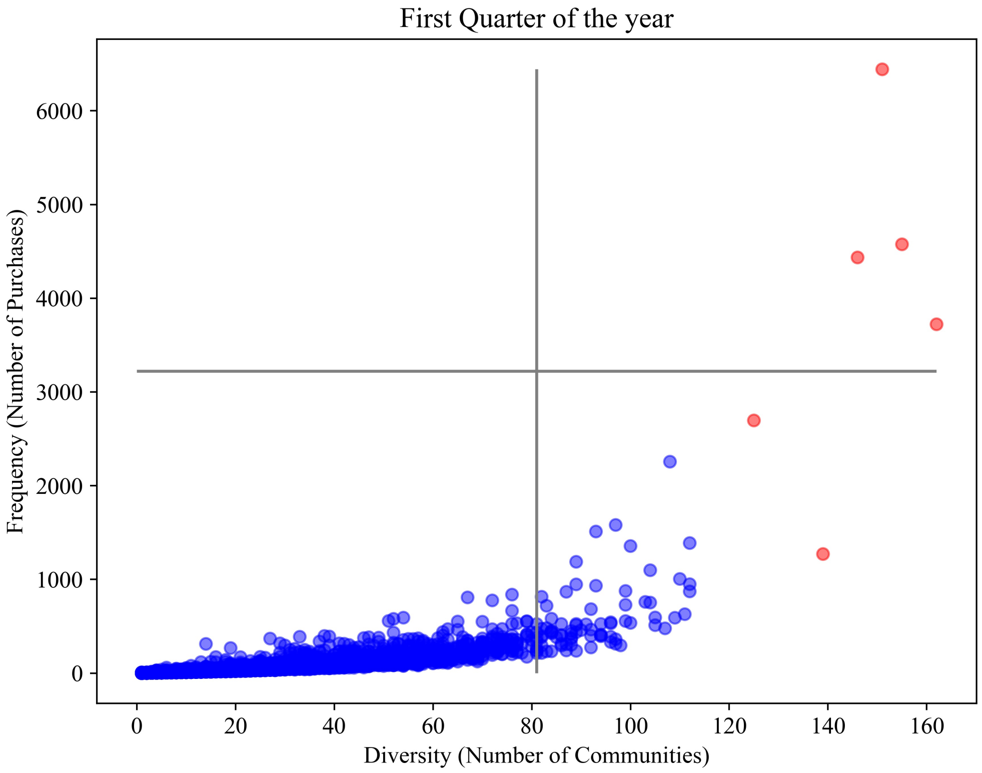

Figure 4.8 Diversity vs. Frequency. Red dots are five customers with the highest diversity score.

Table 4.7 Business implications of the two significant segments of UCI Retail customers.

Customer Segment	First Quadrant (High frequency High diversity, top-right)	Fourth quadrant (Low frequency High diversity, bottom-right)
Business implications	Three customers here are both frequent and diverse customers (shown in red in Figure 4.8). All customers in this segment are targeted in a standard frequency analysis. Frequent customers are generally overwhelmed by inefficient marketing campaigns. This segment helps sharpen the focus of the marketing managers.	The two of the customers shown in red in this segment are under the radar of the RFM model due to low-frequency values. On the other hand, they are the top diverse customers on the platform; thus, they are good candidates that give us an additional shortlist to target when issuing campaigns to increase loyalty.

In all periods of the year in Figure 4.7, the majority of the customers fall in the low frequency, low diversity quadrant. On the other hand, only a few customers appear in the high frequency, low diversity segment. The latter can be regarded as the most valuable customer by the online platform. The RFM analysis can discover this segment as they make the highest number of purchases. However, the low frequency, high diversity segment contains a significant type of customers in terms of business implications. Their purchase pattern is the most diverse, meaning that they make purchases from the highest number of communities. Finally, the last one is the high frequency, low diversity segment, which has customers.

Figure 4.8 shows six customers in red who have the highest diversity score in the first quarter of the year. Three of them are in the first quadrant (top-right), which the standard RFM analysis can easily detect. The two customers in the fourth quadrant (bottom-right) will have relatively low-frequency rankings and be ignored.

However, the diversity dimension of those customers is the highest in the network, indicating that they exhibit very diverse purchase patterns that can be a good target for increasing future purchases with activities such as cross-selling or up-selling. Table 4.7 explains the two significant segments shown in Figure 4.8, the first and fourth quadrants. The empty quadrant is different in this data set. It is the second quadrant this time which contains no customers. The third quadrant contains customers with low values for both attributes presenting no significant business implications in terms of our research; thus, they are excluded from the discussion.

4.11 Customer Segmentation Discussion

This thesis aims to present a novel customer segmentation method based on purchase patterns using community detection algorithms in co-purchase networks extending the RFM analysis. Usually, the RFM model ranks each customer from 1 to 5 for all three parameters, 5 being the highest value. The combination of these three scores is called the customer's RFM cell (Miglautsch 2000). A customer with a 555 cell values, for instance, is the most valuable target for marketers. RFM analysis is based on the Pareto principle, which states that 80% of the business comes from 20% of the customers (Aggelis & Christodoulakis 2005). Therefore, companies focus on detecting this top 20%. This rate is not a strict value; thus, marketers adjust this rate according to their business needs. The diversity score in this research adds a new dimension to the model by quantifying the intricate co-purchase patterns of the customers, helping discover the customers whom the standard RFM model cannot notice. We apply the Pareto principle to the diversity score by focusing on the customers with top diversity scores. The intended outcome is to make a customer return to the store and purchase a new product. Customers tend to buy a particular set of items from one platform and others from a different store (Uusitalo 2001, Yurova et al. 2017) for various reasons such as price, availability, and accessibility. Especially the online marketplaces having a wide range of products and services would like the customers to use their platform for all their needs. Therefore, a customer coming back to buy a product that does not match his or her purchase patterns presents a different significance for the platform managers. The diversity score can detect this purchase pattern change by inspecting the co-purchase network.

A considerable amount of literature has been published on RFM that extends the model with new features to address various issues (Khajvand et al. 2011, Christy et al. 2021, Noori 2015). The new methods are often named by prefixes or suffixes such as RFMD and RFM-N. While some research has mentioned the diversity concept (Hajiha et al. 2011, Burgiel & Sowa 2017, Zhang et al. 2015), one research used the concept to explain the variety of the clustering methods along with the RFM model (Lefait & Kechadi 2010). Diversity term in this thesis refers to the amount of variety in customers' purchase patterns. Although one can employ the method presented in this research by using all three components of the RFM, another option is to use the only most relevant component in combination with the diversity measure. This research ignores the recency and the monetary attributes while combining the frequency attribute with the diversity. Our findings show that the frequency component correlates relatively more with diversity behavior. The below sections discuss the correlation values of each data set in detail. In the meantime, to visually elaborate on the most diverse customers' place in the standard RFM model, Figure 4.9 illustrates the RFM values of the EMP customers and the top ten diverse customers. The red dots represent the top diversity scored customers, and the rest of the customers are depicted in blue. As seen in the figure, customers exhibiting high diversity are on the low ends of the monetary scale. It is worth noticing that high diversity customers are also among the most recent customers (the most recent customer has zero recency value).

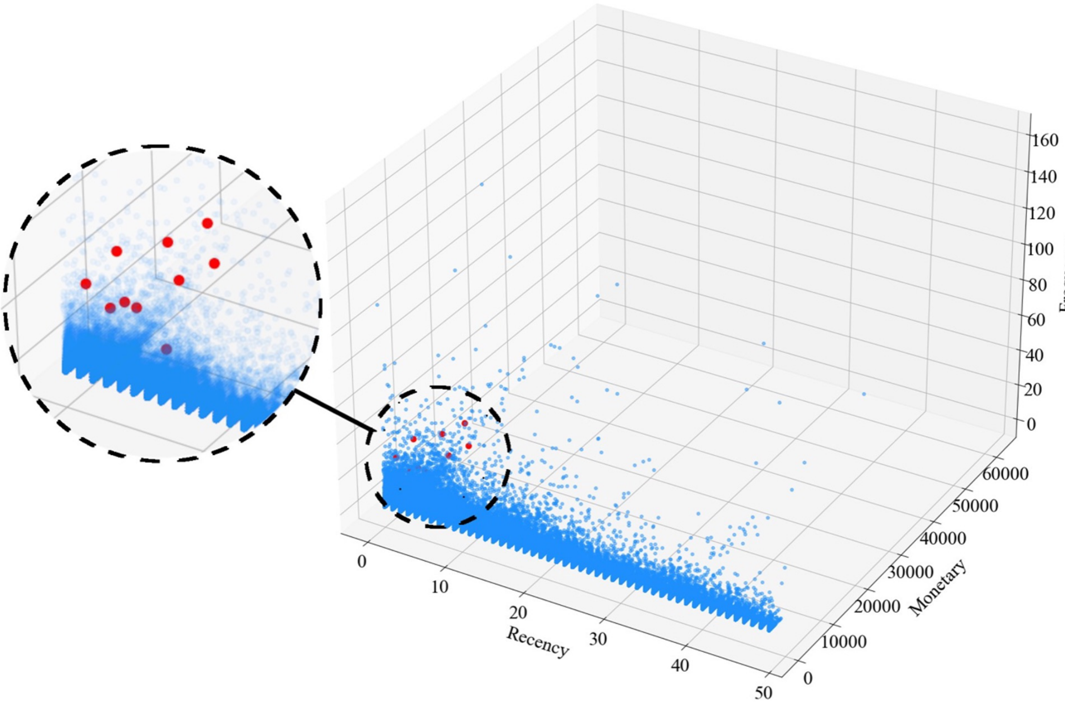

Figure 4.9 RFM values of EMP customers. The top ten diversity-scored customers are in red.

5. INCLUDED PAPERS AND CONTRIBUTIONS

5.1 Paper 1

Perdahçı, Z.N., Aydın, M.N. and Kafkas, K., 2018, Validity Issues in Linked Data Driven IS Research. (Perdahçı et al. 2020)

5.1.1 Summary

This is one of the earlier papers of our work on network analysis research that addresses the validity issues researchers face when conducting Social Network Analysis. Although its scope is SNA studies in general, education domain applications are used to exemplify the validity issues. Prior to addressing these issues, the network concept is briefly explained. Furthermore, a conceptual model is presented which covers Network Science processes and how the linked data advances starting from the real-world system and IS to complex systems and finally analyzed to produce scientific output.

The validation issues mostly arise between the phase transitions. Data reliability at the beginning when deciding the nodes and links is not likely to cause serious validation problems since these entities are well defined in education networks. However, decisions about the link types, weights, or even non-existence of a link are potentially critical validation checkpoints during the SNA process. Another type of validation issue arises due to temporal issues when deciding whether the analysis is static or dynamic. The interpretation of the algorithm results should involve the effects of time over the network. Additionally, as in every scientific research, the utilized tools have validity issues as well as the measures. Researchers should be aware of the strengths and weaknesses of their tools and metrics.

5.1.2 Contributions

To a relatively new discipline as the Network Science, this paper informs the researchers about the pitfalls throughout the SNA processes. This paper does not claim to address all possible issues. Rather, it is intended to be used by researchers as a starting point to avoid these issues and as a validation checklist after their research. To that end, issues are collected and examined in detail; furthermore, practical solutions are offered to facilitate the researchers in their network analysis efforts.

5.2 Paper 2

Kafkas, K., Perdahçi, N.Z. And Aydin, M.N., 2020, Ground Truth And Metadata Relationship In Sbm Community Detection: School Friendship Network. Yönetim Bilişim Sistemleri Dergisi, 6(1), pp.79-85. Kafkas et al. (2019)

5.2.1 Summary

This is the following work that introduces the Stochastic Block Model community detection algorithm to our research. Moreover, In this study, we employ a version of SBM called NeoDCSBM that compares metadata with the ground truth. Many data sets which are studied by Information Systems researchers involve networks that exhibit community structure. Dividing the large networks into manageable groups (communities) is a crucial first step to understanding the network on a macro scale. Which then enables the researchers to analyze the data on a meso-scale. In our previous work, we presented the Stochastic Block Model approach and compared the metadata with the ground truth. In this study, we introduce a statistical technique called neoSBM that can reveal the relationship between metadata and the community structure on the same real-world school best friendship data set.

These findings agree with the previous paper's findings, except that the previous work had a higher resolution. Nevertheless, we see that the friendship network involves a slightly different community structure than class metadata can explain. We can say that the neoDCSBM method can be used to statistically diagnose the relationship between metadata and the ground truth. With this in mind, we need to quantify this relationship with a sound statistical method, and our research group is working on Block Model Entropy

5.2.2 Contributions

In this study, we employ the neoDCSBM algorithm (a degree corrected extension of neoSBM) to find the relationship between metadata and ground truth using a real-world best friendship network and compare the new findings with the previous work. This work is an effort to validate and evaluate the performance of the method by inspecting the relevance of the metadata and the ground truth. Our aim is to present solutions to IS problems with community understanding to establish research capacity for IS community.

5.3 Paper 3

Kafkas, K., PERDAHÇI, Z.N. and AYDIN, M.N., 2021. Ground Truth in Network Communities and Metadata-Aware Community Detection: A Case of School Friendship Network. Alphanumeric Journal, 9(1), pp.49-62. (Kafkas et al. 2021*b*)

5.3.1 Summary

Real-world networks are everywhere and can represent biological, technological, and social interactions. They constitute complicated structures in terms of the type of things and their relations. Understanding the network requires a better examination of the network structure that can be achieved at various scales, including macro, meso, and micro. This research is concerned with the meso scale for a student best friendship network where sub-structures in which groups of entities (students) take different functions. In this study, we address the following research questions: To what extent would NeoSBM as a stochastic process underlie best friendship interaction and, in turn, ground truth interactions (i.e., reported best friendship)? Do metadata such as gender or class contribute to this understanding? How can one support school managers from a meta-data aware community detection perspective? Our findings suggest that metadata aware community detection can be an effective method in supporting decision-making for the class formation and group formation for in and out school activities. Keywords: SBM, neoSBM, Community Detection, Best Friends Network.

5.3.2 Contributions

In this paper, findings agree with the literature, e.g., Perdahcı et al. (2019), except that the previous work had higher resolution with eight communities that divided class 10E and 10D to two subgroups. Nevertheless, we see that the friendship network involves a slightly different community structure than class metadata can explain. One can say that the neoDCSBM method can be used to statistically diagnose the relationship between metadata and the ground truth. With this in mind, we need to quantify this relationship with a sound statistical method, and our research group is working on Blockmodel Entropy Significance Test (BESTest), which computes the entropy of the SBM that describes the detected partitions (Peel et al. 2017).

As for the managerial Implications of the second largest component, Newman (2006) argues that the building blocks are largely invariant with respect to a selected community detection algorithm. If that is the case, investigating the building blocks should be as important, if not more important, as community detection.

5.4 Paper 4

Kafkas, K., Perdahçı, Z.N. and Aydın, M.N., 2021. Discovering Customer Purchase Patterns in Product Communities: An Empirical Study on Co-Purchase Behavior in an Online Marketplace. Journal of Theoretical and Applied Electronic Commerce Research, 16(7), pp.2965-2980. (Kafkas et al. 2021*a*)

5.4.1 Summary

This paper constitutes the product based segmentation half of the thesis. In this research, we build a co-purchase network and empirically study the transaction data of an online platform. 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 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.

5.4.2 Contributions

This study discovered customers' purchase patterns by examining product network communities using 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

5.5 Paper 5

This paper covers the customer segmentation part of this thesis. We submitted the paper to the Journal of Management Science and Engineering. It is waiting for the response of the editors.

6. CONCLUSION

E-commerce has developed into a practical tool for businesses to better serve their clients by integrating online sales and marketing activities. The goal of customer relationship management (CRM) is to collect a large amount of information about customers, such as their purchase history, behavior patterns, and preferences, in order to determine the customers' needs and develop customized recommendations, targeted campaigns, and convincing marketing messages that are likely to be of high relevance to the customers. This results in an increase not only in sales and revenue but also in the level of customer satisfaction and loyalty to the service. With the ever-increasing number of products, services, and customers, businesses are acutely aware of the critical need to segment these entities into smaller groups in order to better understand the massive amounts of digital trace data they have accumulated in order to achieve a competitive advantage over their rivals.

Analysis of purchase behavior involves the two critical components; the product being sold and the customer who makes the purchase. Both of which hold intricate purchase behavior patterns that offer valuable insights to the decision-makers on developing effective marketing strategies. MBA is a conventional method for analyzing the relationship between products, and the RFM technique is traditionally used to segment customers based on their purchase habits. Relatively recently, with the advances in computational capabilities, the network approach has been introduced to the area with models such as link analysis, which involves building a network by linking frequently purchased products together. Researchers have been implementing Social Network Analysis techniques in these co-purchase networks with exciting results for many years. Although not as common as network centrality metrics, researchers apply various community detection algorithms to the co-purchase networks.

This thesis focuses on SBM community detection on co-purchase networks to perform product and customer segmentation. Stochastic Block Modeling is a principled method that makes use of Bayesian statistical inference. The purpose of this research was to uncover the purchasing habits of consumers by analyzing product network communities using SBM. Because it is a probabilistic and generative model, SBM provides a superior solution compared to heuristics-based methods like modularity maximization, which have a tendency to overfit the data and struggle to discover latent communities in large networks. Being a statistically principled method makes SBM's findings independent and reduces the likelihood of them including errors. Therefore, it is not just a scientific invention but also a newly developed scientific technology that is appropriate for use in decision support systems for all forms of electronic commerce. In a very short amount of time, this innovative piece of scientific technology may be included in the preexisting decision-support systems of various online marketplaces. Marketing managers are able to optimize marketing operations such as product suggestion, product placement, cross-selling, and customer retention by segmenting items depending on the purchasing habits of customers and the roles they play in the network.

The stochastic nature of the SBM causes the output to change with each run of the algorithm, with only a few items being allocated to various communities at each run of the algorithm. This presents a challenge for our research since it limits the generalizability of our findings. Within the scope of this investigation, the monetary characteristic served as the basis for the edge weights of the SBM. We were able to notice its impacts in the findings, which showed that the algorithm had a tendency to group financially comparable goods together in the same communities. In future work, either the frequency or recency information may be chosen to examine the results, or all of the alternative edge weights can be utilized to discover which one matches the data the best. Alternatively, the results can be observed regardless of which information was chosen. Moreover, instead of product-to-product networks, constructing seller-to-seller or buyer-to-buyer networks and performing the segmentation methods proposed in this thesis could support marketplace managers in assessing the value of the buyers and sellers in their platforms. The contributions of this thesis are summarized in Table 6.1 and limitations are summarized in Table 6.2.

Table 6.1 Contributions of the thesis.

Product Based Segmentation	**Customer Segmentation**
• Detecting products that play a topologically central role or a bridge role in the co-purchase behavior of the customers.	• A novel metric called diversity is presented in this thesis that quantifies the number of the similar purchase patterns of a customer
• The business implications pertaining to product based segmentation results are described in Table 4.5.	• The business implications pertaining to customer segmentation results are described in Table 4.7.
• A statistically principled method inference method is utilized in this thesis.	• A statistically principled method inference method is utilized in this thesis.
• There is no resolution problem in this methodology. Therefore, it can run on large networks, especially with the hierarchical version.	• There is no resolution problem in this methodology. Therefore, it can run on large networks, especially with the hierarchical version.

In this thesis, empirical research is carried out on two distinct transaction data sets in order to accomplish customer segmentation using a unique measure known as the diversity score. Combining the frequency component of the RFM model with the diversity metric is the method that we propose in order to enhance the segmentation capabilities of the RFM model. In order to compute the level of diversity, we do an analysis of the interactions inside co-purchase networks using principled community identification techniques. According to our research, there is a sizeable population of clients that have high diversity ratings. Even more significantly, they do not get recognized since they fall below the significance threshold in terms of recency, frequency, and monetary values. Because the correlation findings show that the

Table 6.2 Limitations of the thesis.

Product Based Segmentation	Customer Segmentation
• As the “no free lunch” theorem implies (Peel et al. 2017), there is no single community detection method that can work on all networks for all purposes. Therefore, one should utilize a suitable version of the community detection algorithm. • Due to being a probabilistic method, a small number of products may be assigned to a different community after each run.	• Only the frequency component is combined with the RFM model in this thesis. A full model can also be implemented. • As the “no free lunch” theorem implies (Peel et al. 2017), there is no single community detection method that can work on all networks for all purposes. Therefore, one should utilize a suitable version of the community detection algorithm. • Due to being a probabilistic method, a small number of products may be assigned to a different community after each run.

frequency component of the RFM is much more connected to the diversity score, we solely combine that component with the diversity score. The significance of this research may essentially be summed up as having two parts. First, the diversity score that was used in this research brings a new facet to the RFM model, which allows for the identification of clients that are not taken into account by the conventional model. This strategy identifies a new category of consumers who, in contrast to the others, buy a wide variety of items, which, if successful, will ideally contribute to an increase in the number of marketing activities that include cross-selling, up-selling, customer retention, and customer loyalty. Second, we use a statistically principled community identification technique in this investigation to find the hidden product communities. This distinguishes our approach from the heuristic ones that have been previously used. The diversity measure has potential use as a stand-alone methodology for use in further research endeavors. Additionally, a technique that identifies the characteristics of the product communities would be a realistic solution for the decision-makers to better understand the customer segments. This solution would also include determining the characteristics of the product communities.

REFERENCES

Aggelis, V. & Christodoulakis, D. (2005), Customer clustering using rfm analysis, Citeseer, p. 2.

Agrawal, R., Imieliński, T. & Swami, A. (1993), Mining association rules between sets of items in large databases, pp. 207–216.

Agrawal, R. & Srikant, R. (1994), Fast algorithms for mining association rules, Vol. 1215, Citeseer, pp. 487–499.

Aicher, C., Jacobs, A. Z. & Clauset, A. (2015), 'Learning latent block structure in weighted networks', *Journal of Complex Networks* **3**(2), 221–248. Number: 2 Publisher: Oxford University Press.

Akter, S. & Wamba, S. F. (2016), 'Big data analytics in E-commerce: a systematic review and agenda for future research', *Electronic Markets* **26**(2), 173–194. Publisher: Springer.

Ballestar, M. T., Grau-Carles, P. & Sainz, J. (2018), 'Customer segmentation in e-commerce: Applications to the cashback business model', *Journal of Business Research* **88**, 407–414. Publisher: Elsevier.

Barabási, A.-L. (2013), 'Network science', **371**(1987), 20120375. Publisher: The Royal Society Publishing.

Burgiel, A. & Sowa, I. (2017), 'New consumer trends adoption by generations X and Y-comparative analysis', *Zeszyty Naukowe Szkoły Głównej Gospodarstwa Wiejskiego. Ekonomika i Organizacja Gospodarki Żywnościowej* (117).

Büchter, O. & Wirth, R. (1998), Discovery of association rules over ordinal data: A new and faster algorithm and its application to basket analysis, Springer, pp. 36–47.

Chattopadhyay, S., Basu, T., Das, A. K., Ghosh, K. & Murthy, L. C. (2020), 'Towards effective discovery of natural communities in complex networks and implications in e-commerce', *Electronic Commerce Research* pp. 1–38. Publisher: Springer.

Chen, D., Sain, S. L. & Guo, K. (2012), 'Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining', *Journal of Database Marketing & Customer Strategy Management* **19**(3), 197–208. Publisher: Springer.

Christy, A. J., Umamakeswari, A., Priyatharsini, L. & Neyaa, A. (2021), 'RFM ranking–An effective approach to customer segmentation', *Journal of King Saud University-Computer and Information Sciences* **33**(10), 1251–1257. Publisher: Elsevier.

Csardi, G. & Nepusz, T. (2006), 'The igraph software package for complex network research', **1695**(5), 1–9.

Ding, Z., Hosoya, R. & Kamioka, T. (2018), 'Co-Purchase Analysis by Hierarchical Network Structure'.

Esmaeili, L. & Alireza Hashemi Golpayegani (2021), 'A novel method for discovering process based on the network analysis approach in the context of social commerce systems', *Journal of theoretical and applied electronic commerce research* **16**(2), 34–62. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Faridizadeh, S., Abdolvand, N. & Harandi, S. R. (2018), Market basket analysis using community detection approach: A real case, *in* 'Applications of Data Management and Analysis', Springer, pp. 177–198.

Fortunato, S. & Barthelemy, M. (2007), 'Resolution limit in community detection', *Proceedings of the national academy of sciences* **104**(1), 36–41. Number: 1 Publisher: National Acad Sciences.

Freeman, L. C. (1977), 'A set of measures of centrality based on betweenness', *Sociometry* pp. 35–41. Publisher: JSTOR.

Gabardo, A., Berretta, R. & Moscato, P. (2019), Overlapping communities in co-purchasing and social interaction graphs: a memetic approach, *in* 'Business and Consumer Analytics: New Ideas', Springer, pp. 435–466.

Ghasemian, A., Hosseinmardi, H. & Clauset, A. (2019), 'Evaluating overfit and underfit in models of network community structure', *IEEE Transactions on Knowledge and Data Engineering* **32**(9), 1722–1735. Number: 9 Publisher: IEEE.

Guimera, R., Sales-Pardo, M. & Amaral, L. A. N. (n.d.), 'Modularity from fluctuations in random graphs and complex networks', **70**(2), 025101. Publisher: APS.

Hajiha, A., Radfar, R. & Malayeri, S. S. (2011), Data mining application for customer segmentation based on loyalty: An iranian food industry case study, IEEE, pp. 504–508.

Holland, P. W., Laskey, K. B. & Leinhardt, S. (1983), 'Stochastic blockmodels: First steps', *Social networks* **5**(2), 109–137. Number: 2 Publisher: Elsevier.

Huang, Z., Zeng, D. D. & Chen, H. (2007), 'Analyzing consumer-product graphs: Empirical findings and applications in recommender systems', *Management science* **53**(7), 1146–1164. Number: 7 Publisher: INFORMS.

Kafkas, K., Perdahçı, N. Z. & Aydın, M. N. (2019), 'Ground Truth and Metadata relationship in SBM Community Detection: School Friendship Network', **6**(1), 79–85.

Kafkas, K., Perdahçı, Z. N. & Aydın, M. N. (2021*a*), 'Discovering Customer Purchase Patterns in Product Communities: An Empirical Study on Co-Purchase Behavior in an Online Marketplace', *Journal of Theoretical and Applied Electronic Commerce Research* **16**(7), 2965–2980. Publisher: Multidisciplinary Digital Publishing Institute.

Kafkas, K., Perdahçı, Z. N. & Aydın, M. N. (2021*b*), 'Ground truth in network communities and metadata-aware community detection: A case of school friendship network', **9**(1), 49–62.

Kamakura, W. A. (2008), ‘Cross-selling: Offering the right product to the right customer at the right time’, *Journal of Relationship Marketing* **6**(3-4), 41–58. Number: 3-4 Publisher: Taylor & Francis.

Kamakura, W. A., Wedel, M., De Rosa, F. & Mazzon, J. A. (2003), ‘Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction’, *International Journal of Research in marketing* **20**(1), 45–65. Number: 1 Publisher: Elsevier.

Karrer, B. & Newman, M. E. (2011), ‘Stochastic blockmodels and community structure in networks’, *Physical review E* **83**(1), 016107. Number: 1 Publisher: APS.

Khajvand, M., Zolfaghar, K., Ashoori, S. & Alizadeh, S. (2011), ‘Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study’, *Procedia Computer Science* **3**, 57–63. Publisher: Elsevier.

Kim, H. K., Kim, J. K. & Chen, Q. Y. (2012), ‘A product network analysis for extending the market basket analysis’, *Expert Systems with Applications* **39**(8), 7403–7410. Number: 8 Publisher: Elsevier.

Korczak, J., Pondel, M. & Sroka, W. (2019), An approach to customer community discovery, IEEE, pp. 675–683.

Lees, G., Winchester, M. & De Silva, S. (2016), ‘Demographic product segmentation in financial services products in Australia and New Zealand’, *Journal of Financial Services Marketing* **21**(3), 240–250. Number: 3 Publisher: Springer.

Lefait, G. & Kechadi, T. (2010), Customer segmentation architecture based on clustering techniques, IEEE, pp. 243–248.

Liao, S.-H., Chen, Y.-J. & Yang, H.-W. (2013), ‘Mining customer knowledge for channel and product segmentation’, *Applied Artificial Intelligence* **27**(7), 635–655. Number: 7 Publisher: Taylor & Francis.

Ma’arif, M. R. & Mulyanto, A. (2014), ‘Improving Recommender System Based on Item’s Structural Information in Affinity Network’, *Proceeding of the Electrical Engineering Computer Science and Informatics* **1**(1), 186–189. Number: 1.

McCarthy, A. D., Chen, T. & Ebner, S. (2019), An exact no free lunch theorem for community detection, Springer, pp. 176–187.

McKinney, W. (2010), Data structures for statistical computing in python, Vol. 445, Austin, TX, pp. 51–56. Issue: 1.

Miglautsch, J. R. (2000), 'Thoughts on RFM scoring', *Journal of Database Marketing & Customer Strategy Management* **8**(1), 67–72. Publisher: Springer.

Newman, M. (2008), 'The mathematics of networks. The new palgrave encyclopedia of economics'. Publisher: Palgrave Macmillan Basingstoke.

Newman, M. E. (2006), 'Modularity and community structure in networks', *Proceedings of the national academy of sciences* **103**(23), 8577–8582. Number: 23 Publisher: National Acad Sciences.

Noori, B. (2015), 'An Analysis of Mobile Banking User Behavior Using Customer Segmentation.', *International Journal of Global Business* **8**(2).

Oestreicher-Singer, G., Libai, B., Sivan, L., Carmi, E. & Yassin, O. (2013), 'The network value of products', *Journal of Marketing* **77**(3), 1–14. Number: 3 Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Peel, L., Larremore, D. B. & Clauset, A. (2017), 'The ground truth about metadata and community detection in networks', *Science advances* **3**(5), e1602548. Number: 5 Publisher: American Association for the Advancement of Science.

Peixoto, T. P. (2014*a*), 'The graph-tool python library. figshare', **10**, m9.

Peixoto, T. P. (2014*b*), 'Hierarchical block structures and high-resolution model selection in large networks', *Physical Review X* **4**(1), 011047. Number: 1 Publisher: APS.

Peixoto, T. P. (2018), 'Nonparametric weighted stochastic block models', *Physical Review E* **97**(1), 012306. Number: 1 Publisher: APS.

Peixoto, T. P. (2019), 'Bayesian stochastic blockmodeling', *Advances in network clustering and blockmodeling* pp. 289–332. Publisher: Wiley Online Library.

Peixoto, T. P. (2020), 'Merge-split Markov chain Monte Carlo for community detection', *Physical Review E* **102**(1), 012305. Number: 1 Publisher: APS.

Perdahcı, Z. N., Aydın, M. N. & Kafkas, K. (2019), 'SBM based community detection: School friendship network'. Publisher: IMISC.

Perdahçı, Z. N., Aydın, M. N. & Kafkas, K. (2020), 'Validity issues in linked data driven IS research'.

Puka, R. & Jedrusik, S. (2021), 'A New Measure of Complementarity in Market Basket Data', *Journal of Theoretical and Applied Electronic Commerce Research* **16**(4), 670–681. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Raeder, T. & Chawla, N. V. (2009), Modeling a store's product space as a social network, IEEE, pp. 164–169.

Shi-Yong, Z., Mao-Hong, L. & Jin-De, H. (2019), 'The Influence of Community Structure on the Diffusion of Knowledge–A View Based on Market Segmentation.', *International Journal of Emerging Technologies in Learning* **14**(8). Number: 8.

Suryateja, G. & Palani, S. (2017), Survey on efficient community detection in social networks, IEEE, pp. 93–97.

Tsiptsis, K. K. & Chorianopoulos, A. (2011), *Data mining techniques in CRM: inside customer segmentation*, John Wiley & Sons.

Uusitalo, O. (2001), 'Consumer perceptions of grocery retail formats and brands', *International Journal of Retail & Distribution Management* . Publisher: MCB UP Ltd.

Videla-Cavieres, I. F. & Rios, S. A. (2014), 'Extending market basket analysis with graph mining techniques: A real case', *Expert Systems with Applications* **41**(4), 1928–1936. Number: 4 Publisher: Elsevier.

Vindevogel, B., Van den Poel, D. & Wets, G. (2005), 'Why promotion strategies based on market basket analysis do not work', *Expert Systems with Applications* **28**(3), 583–590. Number: 3 Publisher: Elsevier.

Wang, D., Li, J., Xu, K. & Wu, Y. (2017), 'Sentiment community detection: exploring sentiments and relationships in social networks', *Electronic Commerce Research* **17**(1), 103–132. Number: 1 Publisher: Springer.

Wang, S.-C., Hsu, H.-W., Dai, C.-G., Ho, C.-L. & Zhang, F.-Y. (2019), Use Product Segmentation to Enhance the Competitiveness of Enterprises in the IoT, IEEE, pp. 1–6.

Woo, J. (2013), 'Market basket analysis algorithms with mapreduce', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **3**(6), 445–452. Number: 6 Publisher: Wiley Online Library.

Yurova, Y., Rippé, C. B., Weisfeld-Spolter, S., Sussan, F. & Arndt, A. (2017), 'Not all adaptive selling to omni-consumers is influential: The moderating effect of product type', *Journal of Retailing and Consumer Services* **34**, 271–277. Publisher: Elsevier.

Zhang, L., Priestley, J., DeMaio, J., Ni, S. & Tian, X. (2021), 'Measuring Customer Similarity and Identifying Cross-Selling Products by Community Detection', *Big Data* **9**(2), 132–143. Number: 2 Publisher: Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor.

Zhang, Y., Bradlow, E. T. & Small, D. S. (2015), 'Predicting customer value using clumpiness: From RFM to RFMC', *Marketing Science* **34**(2), 195–208. Publisher: INFORMS.

APPENDIX A: PARAMETER SELECTION

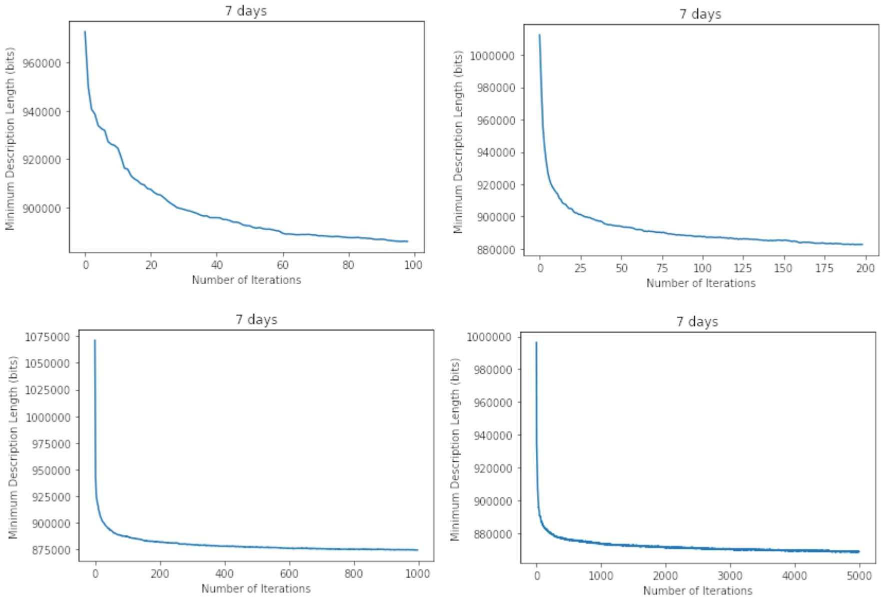

Figure A.1 The elbow plots of iteration parameter selection for seven days EMP co-purchase network. 100, 200, 1000, and 5000 iterations from top-left to bottom-right.

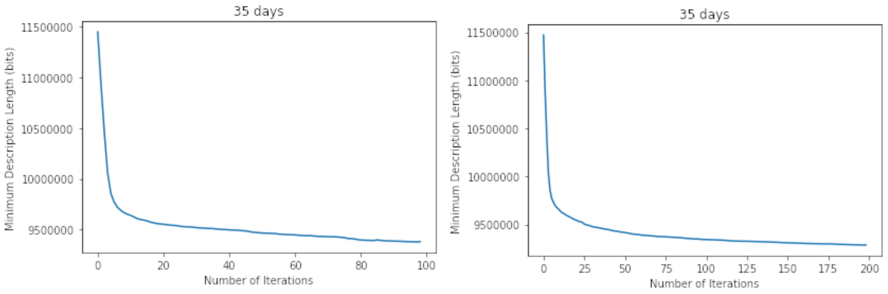

Figure A.2 The elbow plots of iteration parameter selection for 35 days EMP co-purchase network. 100, 200 iterations.

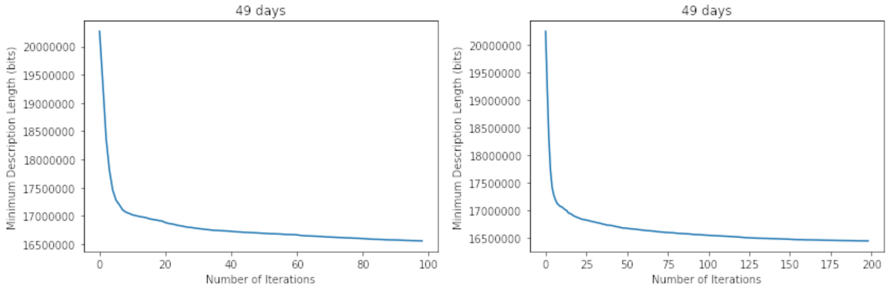

Figure A.3 The elbow plots of iteration parameter selection for 49 days EMP co-purchase network. 100, 200 iterations.

CIRRICULUM VITAE

Personal Information

Name and surname	: Kenan Kafkas

Academic Background

Bachelor's Degree Education:	Marmara University Faculty of Science and Literature B.Sc. in Mathematics
Graduate Education	: Kadir Has University M.Sc. In Management Information Systems

Work Experience

Institutions and Dates	: Ministry of National Education High School Mathematics Teacher 2002 - Present