NORTHWESTERN UNIVERSITY

# Multi-Stage Customer Preferences Modeling Using Data-Driven Network Analysis

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Mechanical Engineering

By

Yaxin Cui

EVANSTON, ILLINOIS

September 2023

# ABSTRACT

This dissertation aims to develop innovative analytical methods that integrate engineering, marketing, and social science disciplines to incorporate heterogeneous consumer preferences into product design using network-based customer preference modeling. Both companies and designers frequently face difficulties in understanding and addressing customer preferences, which can result in product failure and loss of market share. To overcome the limitations of existing methodologies, this dissertation presents a novel approach emphasizing network-based methods for modeling and analyzing customer preferences in engineering design and market research. By representing relations between customers and products as intricate networks and utilizing data-driven network analysis, this approach facilitates a deeper understanding of customer preferences. Consequently, it enhances product design and marketing strategies by effectively employing network-based techniques for preference modeling.

The proposed approach comprises several key methodological developments, all focused on the concept of network-based customer preference modeling. First, a weighted network modeling approach for product competition analysis that captures the competition strength is introduced. This approach utilizes weighted network modeling and predictive analytics to examine product competition. Importantly, by quantifying the link strength in the network, this method offers a detailed understanding of the competitive landscape, identifying factors contributing to product success or failure in the market.

Second, a framework incorporating information retrieval and survey design is developed to investigate customers' two-stage decision-making processes and the influence of their social networks. This framework captures the intricacies of consumer preferences and decision-making, revealing how preferences differ in the consideration and choice stages, and how social influence

affects these preferences. The effectiveness of the proposed approach is demonstrated through a case study on household vacuum cleaners, highlighting its ability to capture consumer preferences and guide product design decisions.

Third, a network-based analysis of heterogeneous customer preference modeling with market segmentation is presented. The proposed techniques enable an examination of varied customer preferences, aiding businesses in creating products that appeal to distinct market segments. Understanding the hierarchical structure of preferences and the underlying decision-making processes enables companies to customize marketing strategies effectively, targeting specific customer groups.

Lastly, graph neural network-based methods in Link Prediction are investigated, concentrating on unidimensional product competition networks and preliminary findings on bipartite customer consideration-then-choice networks. These methods exhibit the potential for predicting consumer preferences and choices, providing valuable insights for both product design and marketing strategies. By integrating these advanced machine learning techniques, the proposed approach demonstrates its capacity to reveal complex patterns in consumer preferences, leading to a more comprehensive understanding of customer behavior.

The proposed methodology equips engineering designers with the methodology and tools to better understand and respond to customer preferences and market trends, leading to more effective product design and marketing strategies. The contributions of this research have significant implications for both academia and industry, particularly in improving the design and marketing of consumer products. By employing network-based customer preference modeling, this dissertation offers an innovative approach to comprehending the complex nature of consumer preferences and their influence on product success. By highlighting the significance of network-based customer preference modeling and demonstrating its effectiveness through various contributions, this dissertation lays a robust foundation for future work in this area. Expanding these analytical methods to other industries and exploring additional network-based techniques will further enhance our understanding of consumer preferences, driving the development of successful products that cater to diverse customer needs. As the field continues to evolve, the insights gained from this research will play a crucial role in shaping the future of product design and marketing strategies.

## ACKNOWLEDGEMENTS

First and foremost, I wish to convey my deepest appreciation to my advisor, Prof. Wei Chen, whose steadfast support and mentorship were pivotal throughout my doctoral journey. Her wisdom, rigorous critiques, and unfaltering guidance have proven indispensable to my academic and personal growth. Her patience and faith in my abilities have been a constant source of motivation, encouraging me to strive towards excellence.

I would also like to express my profound gratitude to my collaborators, Prof. Zhenghui Sha, Prof. Noshir Contractor, and Prof. Johan Koskinen, as well as my committee member, Prof. Elizabeth Gerber. Their insightful suggestions and invaluable feedback have greatly enriched my work. My sincere appreciation also extends to my fellow Postdoc and Ph.D. students from our collaborative lab, especially Yinshaung Xiao, Neelam Jignesh Modi and Faez Ahmed. Their camaraderie, thought-provoking discussions, and unique perspectives significantly contributed to my research and were a constant source of inspiration. I am deeply grateful for their assistance in fostering my critical thinking and enabling me to explore new facets of my research.

Special acknowledgment is owed to my peers in the IDEAL lab. The stimulating environment of intellectual curiosity, mutual respect, and shared passion for research that we cultivated has been instrumental in my journey. I am particularly grateful to Zhuoxin (Joy) Sun, whose assistance during the stressful graduation season was invaluable. Her expertise in data analytics, writing, and critical thinking have been pivotal in the completion of my final thesis and papers. The IDEAL lab's camaraderie has fostered a nurturing environment for learning, growth, and mutual support, for which I am profoundly thankful.

I am deeply appreciative of my friends, many of whom have accompanied me throughout my doctoral journey. Special thanks go to Dingwen Qian, Wei Wang, Mengfan Xu, and Yi Wang

and all others who have provided me with constant moral support and encouragement. Their presence and friendship have been a tremendous source of strength for me. I am also profoundly grateful to my friends from the Northwestern Sheil Catholic Center for their spiritual support and encouragement in all my endeavors.

I am eternally indebted to my family for their unconditional love, unwavering support, and belief in my abilities. My parents have been the bedrock of my strength, continually providing encouragement and solace during challenging times. Their faith in my dreams has led them to make considerable sacrifices, including not meeting with me for four years. Their support has been an invaluable part of my journey. I also owe a heartfelt debt of gratitude to my cousin, Xufei Wang, who has been my roommate for the past two years. Her support extends beyond just being a part of my life. She has been an insightful confidante, offering valuable perspectives to overcome challenges in my work. Her contributions have made my journey smoother and more manageable.

I wish to express heartfelt appreciation to the funding agencies, namely NSF CMMI #2005661 and #2203080, as well as the Intersection Science Fellowship at The Northwestern Institute on Complex Systems. Their invaluable financial support greatly facilitated my research, providing me with crucial resources that were instrumental in completing this work.

Lastly, I want to acknowledge the indispensable role I have played in this journey. I am proud of the resilience, dedication, and perseverance I have demonstrated in this journey, and I believe it's essential to recognize the personal growth and intellectual development I have undergone.

In conclusion, I extend my sincere gratitude to everyone who has been a part of my Ph.D. journey. Your contributions, in numerous and varied ways, have made this achievement possible. Thank you for being part of this remarkable journey.

# Nomenclature

## Symbols

| | |
|---|---|
| $A$ | Customer-desired Attributes |
| $S$ | Respondent background information |
| $U_i$ | Decision maker's utility of selecting alternative i |
| $V_i$ | Observed Utility |
| $\epsilon_i$ | Unobserved utility |
| $P_i$ | The probability of choosing alternative $i$ |
| $G(N, E)$ | Mathematical graph with node set N and edge set E |
| $D_{undirected}$, $D_{directed}$ | Network density |
| $C_i$ | Local clustering coefficient |
| $\mathbf{Y}$ | Random variable of a network |
| $\mathbf{y}$ | Instantiation of a network |
| $\theta$ | Parameter vector indicating the effects of network statistics in ERGM |
| $\mathbf{g}(\mathbf{y})$ | A vector of network statistics in y in ERGM |
| $\kappa(\theta)$ | Normalizing constant in ERGM |

## Acronyms

| | |
|---|---|
| DCA | Discrete Choice Analysis |
| DBD | Decision-Based Design |
| MCPN | Multidimensional Customer-Product Network |
| ERGM | Exponential Random Graph Models |
| GNN | Graph Neural Networks |
| SP | Stated Preferences |
| RP | Revealed Preferences |
| GSN | General Social Networks |
| PSN | Product-specific Social Networks |
| JCA | Joint Correspondence Analysis |
| SVD | Singular Vector Decomposition |
| MCMC | Monte Carlo Markov Chain |
| ML | Machine Learning |

## TABLE OF CONTENTS

## LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1
# PROBLEM DESCRIPTION AND RESEARCH OBJECTIVE

This dissertation is driven by the need to create analytical models that explore customer preferences by integrating knowledge from engineering, marketing, and social science domains, with the ultimate goal of identifying optimal product designs for targeted customer groups. Traditional product design is primarily driven by engineering performance and economic constraints; however, customer choices are influenced by various other factors, such as customer heterogeneity and product competition. To make well-informed design decisions, it is imperative to take into account not only these traditional factors but also customer heterogeneity and market competition. In this study, we will demonstrate the effectiveness of our proposed data-driven network analysis methods of multi-stage customer preference modeling. By accounting for the intricate relationships within a design ecosystem, these methods are proven to be more efficient than conventional approaches.

## 1.1 Background of the study

The purpose of **customer preference modeling** is to understand how customers weigh and make trade-offs among different attributes when making decisions about a product. Incorporating customer preference modeling into engineering design bridges the gap between market research and engineering research. The former focuses on customer preferences, behaviors, and market trends, while the later emphasizes the technical aspects of products. By considering customer preferences as a function of customer attributes and product attributes, customer preference models enable a more comprehensive and effective product design process. These models play a vital role in various aspects of engineering design, such as design attribute selection (to identify key product features appealing to customers) (Hoyle et al., 2009), usage and social context-based design (to align products with customers' practical needs and social expectations) (L. He et al., 2012), product configuration (optimizing feature combinations for targeted customer segments) (Sha, Saeger, et al., 2017), and design of engineering systems (ensuring products meet both technical and customer requirements) (Kumar, Hoyle, et al., 2009; Michalek et al., 2006; Sha & Panchal, 2014).

Existing analytical preference models primarily encompass value-based models and agent-based models. Value-based models are more widely used than agent-based models due to their versatility and ability to accommodate various customer preferences and product attributes. At the core of value-based models is the concept of random utility theory, which suggests that customers make choices based on their perceived utility of each option, with a random component accounting for unobservable factors. These models have attracted significant attention over the past decade, resulting in a wealth of related research within the design community. Among value-based choice models, disaggregate quantitative approaches, such as Discrete Choice Analysis (DCA) (M. E. Ben-Akiva & Lerman, 1985; Sha et al., 2019; K. Train, 1986) and conjoint analysis (Tovares et al., 2013), have been rigorously investigated. These approaches enable researchers to better understand customer preferences and have been applied in various contexts. An example of such application is the enterprise-driven Decision-Based Design (DBD) framework (W. Chen et al., 2013; Wassenaar & Chen, 2003; Wassenaar et al., 2005), which utilizes DCA to estimate the demand for a product, which is determined through the aggregation of customers' choices (W. Chen et al., 2013). This information is subsequently used to assess both production costs and design revenue. To enhance the accuracy of these models, Frischknecht et al. (2010) proposed various metrics to compare DCA models with traditional logistic regression. Still, value-based choice analysis has other limitations (M. Wang, Chen, Huang, et al., 2016), including difficulty in modeling interdependencies among customers, accounting for irrational choice behaviors that are caused by social influence, and addressing the correlation of decisions. These limitations and potential improvements are further discussed in Chapter 2.

**Network analysis** has become a prominent method for the statistical analysis of engineering systems across various domains, including scientific, social, and engineering fields (Albert et al., 2000; Braha et al., 2006; Holling, 2001; M. E. Newman, 2003; Simon, 1977; Wasserman & Faust, 1994). More recently, this approach has been adopted to customer preference modeling to enhance our understanding of customer-product relationships (W. Chen et al., 2020). The fundamental rationale for using a network-based approach in this context is the idea that customer-product relationships can be viewed as complex socio-technical systems, which share similarities with other engineering systems that exhibit dynamic, uncertain, and emerging behaviors. By employing social network theory and techniques, it becomes possible to analyze these systems effectively. Through the identification of structural and topological characteristics in customer-product networks, researchers can uncover patterns embedded in customer-product relationships and model the inherent heterogeneities of both customers and products.

In previous research, a generalized multidimensional customer-product network framework (MCPN) has been proposed to model customer preferences in engineering design (M. Wang, Chen, Huang, et al., 2016). MCPN comprises two distinct layers: one for "customers" and another for "products", as shown in the in Figure 1.1. The product layer encompasses a collection of engineering products P, each characterized by product attributes A and connected based on diverse product association relationships. These relationships can be either directed (e.g., one vehicle is chosen over another) or non-directed (e.g., vehicles are co-considered by customers). The customer layer consists of a population of customers C, with each customer defined by their attributes S. Connections between customers signify social relations or interactions, such as friendships or communication. Customer-product relations between the two layers indicate various human activities (decisions), including consideration (dashed line) and choice (solid line). Multidimensional network analyses can provide a comprehensive examination of all three types of relationships within a system - between customers, between products, and between customers and products - resulting in a deeper understanding of the complex interactions and dynamics at play.

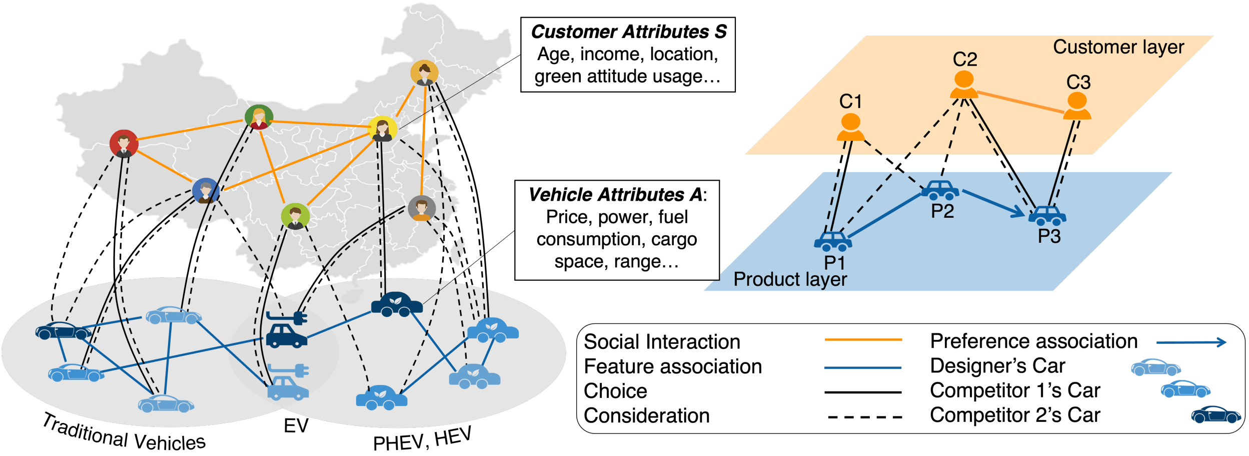


Figure 1.1: Conceptual Framework of Multidimensional Customer-Product Network (MCPN)

To improve our understanding of customer preferences, we must replicate the actual customer decision-making process as closely as possible. The **two-stage (consideration-then-choice) model**, a prominent method in marketing and customer behavior research, posits that customers' decision-making process consists of two stages, consideration and choice, as depicted in Figure 1.1. During the first stage, customers evaluate a set of potential options or brands, forming a consideration set. This set comprises alternatives that could potentially fulfill their needs or desires. Subsequently, in the second stage, customers select an option from the consideration set. In other words, the first stage facilitates the assessment and comparison of a more focused range of alternatives, ultimately leading to a final decision. Understanding customers' two-stage decision-making process is of paramount importance and serves as the central focus of this study. Ultimately, the goal is to provide valuable insights into customer preferences.

## 1.2 Problem statement

As previously discussed, customer preference modeling plays a crucial role in the engineering design field, and a myriad of methods have been developed to capture these preferences. Among these approaches, network-based methods have gained prominence in recent years, with various aspects being explored such as product competition relations (Sha, Huang, Fu, et al., 2018; M. Wang & Chen, 2015), two-stage customer-product relations (Bi et al., 2021; J. Fu et al., 2017), and dynamic network evolution (Xie et al., 2020). Despite the progress and potential of network-based methods in understanding customer preferences, several limitations persist in the existing literature. The following sections will delve deeper into these limitations and outline the research gaps that need to be addressed to advance the field.

**Capturing link weights:** Traditional network statistical methods have reduced networks to binary forms, in which links are either present or absent. This simplification fails to account for the varying strengths of relationships among nodes. For instance, in product competition networks where nodes represent products and links denote their competitive relationships, the 'strength' of a link could be viewed as the intensity of the competition between two products. By neglecting the variations in this competition strength and simply treating all links as binary — present or absent — we may lose valuable information, thereby affecting the accuracy of the resulting models.

**Data availability and systematic data collection:** Data availability and systematic data collection have been major concerns in previous research on network-based customer preference modeling. Specifically, comprehensive data is limited, with the car survey data from the Chinese market being the primary source. In most cases, customer data from other product markets only includes product and customer attributes, along with customers' final choices. Consequently, it becomes difficult to investigate the two-stage decision-making process of customers and the influence of social factors on their decision-making in other markets. Moreover, the lack of data significantly hampers the generalizability of network-based models and restricts the ability to validate research hypotheses across different product communities. To overcome this constraint, a more comprehensive approach to data collection is required that considers product attributes, customer demographics, and decision-making processes simultaneously.

**Addressing customer heterogeneity:** Until now most studies have not sufficiently investigated how network-based methods can be adapted to meet the challenges posed by highly heterogeneous markets. Although these methods have shown promise in modeling complex systems and understanding customer behavior, their effectiveness in dealing with diverse customer preferences, needs, and behaviors in heterogeneous markets remains underexplored. Research is necessary to explore the adaptation and application of market-segmentation-based methods in network-based models in order to better understand and address the challenges posed by highly heterogeneous markets.

**Improving overall model accuracy and complexity:** Existing statistical network-modeling methods suffer from two key limitations. First, they cannot manage a significant number of network structural effects, product attributes, and customer attributes simultaneously. Second, these methods often rely on fixed, linear mathematical forms, making it difficult to capture intricate patterns in the data. As a result, these methods usually have low accuracy when used to predict customers' behaviors. Therefore limits its application to practical problems. Therefore, there is a research need to explore innovative and flexible network-based modeling approaches capable of effectively accommodating larger sets of attributes and network structural effects, thereby improving the accuracy of customer behavior predictions across a variety of market conditions.

## 1.3 Research questions and objectives

The limitations of current network-based analysis methods have prompted a set of research questions that this dissertation aims to address:

- Research question 1: How can we capture and incorporate the strength of product competition links, derived from aggregated customer considerations and choices, into the product competition network model to improve the predictability of product competition in the market?
- Research question 2: How should we collect the data to support customer preference modeling through survey design and information retrieval, ensuring the inclusion of necessary information for analyzing social influence within a two-stage customer decision-making process?
- Research question 3: In a market characterized by diverse customer preferences, what strategies can be employed to partition customers into distinct segments effectively, and how can these market segmentation methods be seamlessly integrated with network-based approaches?
- Research question 4: In the context of network-based modeling, how to effectively incorporate high dimensional customer and product attributes, as well as complex network structures, with the goal of improving the overall accuracy of the model?

The main objective of this dissertation is to enhance the effectiveness of network-based methodologies used in modeling customer preferences by addressing limitations identified in previous research. The study places a particular emphasis on the two-stage decision-making process of customers, which is a critical factor that interlinks various network-based methods explored in this study.

While network-based approaches have been predominantly applied to automotive markets and designs, we aim to extend the applicability of these methods to other product markets. To achieve this, we develop a systematic customer survey design protocol that facilitates the collection of relevant information, enabling us to apply network-based models to a wider range of products and markets.

By addressing the shortcomings of previous research and considering the two-stage decision-making process, this research seeks to enhance the network-based approach for customer preference modeling. Additionally, the aim of this work is to provide models that are more precise, thereby improving the overall performance of the model. Ultimately, this research will contribute to a better understanding of how network-based methods can be applied in modeling customer preferences, providing insights that can be used to inform product design and marketing strategies.

## 1.4 Significance of the study

The significance of this study is rooted in its endeavor to enhance the efficacy of network-based methodologies in modeling customer preferences, particularly for their application in engineering design. By addressing the limitations of prior research and emphasizing the two-stage decision-making process undertaken by customers, this study aims to develop a more precise and effective customer preference modeling approach that can be integrated with product design.

First, this study enhances binary link modeling in network analysis by incorporating link strength, improving the accuracy of product competition modeling, and laying the foundation for link-strength-aware network analysis. Second, this study provides engineering designers with a systematic and comprehensive consider-then-choose customer survey design protocol that can be tailored to a variety of product markets, extending the applicability of these methodologies beyond the automotive industry. While network-based models are beneficial, their use is limited without evaluations across diverse product markets. This research, therefore, serves as a bridge, extending the reach of network-based models to versatile problems. Third, this study rigorously examines the effectiveness of market segmentation-based methods in addressing heterogeneous preferences. Our findings reveal that these approaches effectively capture the unique characteristics of the clearly identifiable sub-markets, ultimately yielding more accurate models of customer preferences. Finally, this study demonstrates the effectiveness of graph neural network-based methods in capturing a broader spectrum of customer and product attributes. By implicitly accounting for the intricate complexities inherent in both unidimensional and bipartite networks, these methods demonstrate a significant advantage over conventional network statistical approaches, thereby resulting in predictions with greater accuracy.

Overall, this study enhances our understanding of network-based methodologies for modeling customer preferences and lays the groundwork for developing more effective and accurate methods to design products that better meet customer needs. These findings have significant implications in the engineering design field and can be used to inform product design and marketing strategies.

## 1.5 Thesis structure

The outline of this dissertation is as follows. Chapter 2 presents the literature review and technical background underlying the research tasks. To address Research Question 1, Chapter 3 proposes a weighted network model approach that captures the product competition strength and examines its effectiveness. Chapter 4 presents a systematic information retrieval and survey design process to address Research Question 2. Data from the vacuum cleaner sector is collected and analyzed in this thesis to provide insights into another product market. In Chapter 5, the effectiveness of network-based methods in analyzing a heterogeneous market with diverse customer preferences is examined, and market segmentation methods are integrated to address Research Question 3. Chapter 6 explores the effectiveness of deep-learning-based models, specifically graph neural network-based methods, in capturing more data information and implicit network structures. This chapter addresses Research Question 4. Finally, Chapter 7 summarizes the contribution of this research and suggests areas for future work.

# CHAPTER 2
# LITERATURE REVIEW AND TECHNICAL BACKGROUND

This chapter provides a comprehensive review of the literature and technical background related to our research on multi-stage customer preference modeling using data-driven network analysis. This chapter is divided into three main sections. Section 2.1 focuses on the various stages of modeling customer preferences in engineering design, including data collection, preference modeling approaches, and multi-stage models such as consideration-then-choice frameworks. Section 2.2 delves into data-driven network analysis, discussing its importance and applications in capturing complex relationships among product features, customer preferences, and market dynamics. Finally, Section 2.3 examines the analytical methods used in network analysis, including statistical network models like Exponential Random Graph Models (ERGM) and deep learning-based models such as Graph Neural Networks (GNN). By presenting a thorough review of these topics, we aim to establish the foundation for our research and highlight the key contributions of our work in the context of existing literature.

## 2.1 Customer preference modeling in engineering design

As we introduced in chapter 1, customer preferences modeling plays a crucial role in engineering design. The primary challenges within this field are data collection and the development of effective modeling techniques. In this section, we delve into the existing methods and approaches used in both areas. Furthermore, we discuss two-stage models as a specific case in preference modeling, highlighting their unique features and benefits.

### 2.1.1 Data collection

Two primary types of data utilized for demand modeling include stated preference (SP) data (Louviere et al., 2000) and revealed preference (RP) data. Revealed preference (RP) involves actual, verifiable choices, such as a customer purchasing a product in reality. In contrast, stated preference (SP) data is typically obtained through controlled choice experiments where respondents indicate their hypothetical purchase intentions. Surveys are commonly employed for collecting SP data to ascertain how individuals may react to various products or features.

Stated choice surveys require respondents to select an option from a choice set, which closely resembles real-life purchase decisions. Choice sets contain several competing alternatives, including a ”survey alternative” (i.e., a new product or an alternative with an improved design), one or more competitor alternatives, and occasionally a ”no choice” option. Alternatives are characterized by customer-desired attributes (A) such as price and warranty, and choice sets can be generated using experimental design techniques. Survey results (choice data) are documented, along with respondent background information (S) including age, income, and product usage. SP data is frequently applied in conjoint analysis-based modeling in the marketing and transportation research literature, which encompasses the analysis of three types of consumer preference data: ratings, rankings, and choice data (M. Ben-Akiva et al., 1992; Bradley & Lang, 1994; Haaijer et al., 1998; Louviere et al., 1993). On the other hand, RP data is commonly associated with discrete choice analysis methods, which are frequently applied in transportation and economic studies.

To address challenges such as respondent fatigue in lengthy surveys and inefficiencies in traditional data collection methods, recent advancements have emerged for the collection and analysis of Stated Preference (SP) and Revealed Preference (RP) data. Researchers have explored survey design issues for optimal preference modeling data collection. Hoyle et al. (2009) devised an algorithm to determine the most suitable design for human appraisal experiments, mitigating respondent fatigue. H. Q. Chen et al. (2012) suggested an approach akin to efficient Global optimization, which reduces survey length by formulating questions based on previous responses. Akai et al. (2010) introduced a query algorithm to update user preference models during data collection, enabling shorter surveys by querying earlier users with analogous preference structures about their preferred product designs.

In addition, big data has become increasingly vital for product improvement, redesign, and innovation (Sawhney et al., 2005). This shift necessitates the development of novel technologies to assimilate, analyze, visualize, and utilize the burgeoning volume of big data. Although it can be challenging to access market survey data on customer consideration sets and choices, open data sources (Parraguez, Maier, et al., 2017) have provided additional opportunities for research in engineering design. The proliferation of Web 2.0 has resulted in vast amounts of information shared on social media platforms, such as forums, blogs, and product reviews websites. Capitalizing on this online presence, crowdsourced design (Gerth et al., 2012) has been introduced, allowing customers to provide direct evaluations of perceptual design attributes. Recent research has also investigated the potential of online customer reviews and opinions to aid engineering design through product design feature detection (Rai, 2012) and product design selection (Z. Wang et al., 2011). Various machine learning approaches have been examined to mine transactional data for concealed purchasing patterns. These include data mining techniques for generating new choice modeling scenarios (M. Wang, Chen, Huang, et al., 2016), assessing the feasibility of Twitter as a product opinion source (Stone & Choi, 2013), producing highly accurate preference predictions (Burnap et al., 2016), and creating market segments from online reviews focused on specific product attributes while identifying attribute importance rankings (Rai, 2012).

Furthermore, instead of considering SP and RP as competing valuation techniques, analysts have started to perceive them as complementary, utilizing the strengths of each type to deliver more precise and potentially more accurate models. This approach is commonly referred to as data enrichment or model fusion in the literature (Mark & Swait, 2004; Merino-Castello, 2003).

In addition to collecting data on stated or revealed preferences, product key features, attributes, and functionalities also need to be treated as explanatory variables in preference modeling.Van Horn et al. (2012) broadened the concept of design analytics, illustrating the effective application of information-to-knowledge transformations through data analytics at each design stage, emphasizing the importance of product features, attributes, and functionalities. Moreover, accounting for customer heterogeneity in demand modeling necessitates the incorporation of customer attributes, such as demographic characteristics, usage context, and personal viewpoints, as input variables. In this context, Tucker and Kim (2011) introduced the preference trend mining algorithm, which employs data mining techniques to analyze customer attributes and preferences, detecting unobservable trends related to product features and attributes, and enabling design engineers to predict the next generation of product functionalities.

### 2.1.2 Preference modeling approaches

The early development of analytical models for customer preference can be traced back to market research, wherein various analytical methods, such as Multiple Discriminant Analysis (R. Johnson, 2011), Factor Analysis (Gorsuch, 1983), Multidimensional Scaling (Green, 1970), Conjoint Analysis (Green, 1970; Green & Krieger, 1991; Green & Srinivasan, 1990; Green & Srinivasan, 1978; Green & Wind, 1975) and Discrete Choice Analysis (DCA) (M. E. Ben-Akiva & Lerman, 1985; K. Train, 1986) were developed. Customer preference modeling methods can be broadly classified into two categories: disaggregate approaches, which utilize individual customer data, and aggregate approaches, which rely on group averages and model market share based on product features and customer group socio-demographic attributes. Compared to aggregate approaches, disaggregate methods offer valuable insights into individual decision-making processes influenced by personal preferences, enabling a more in-depth understanding of the various factors affecting customers' choice behaviors, including individual characteristics and product attributes.

In addition to the aforementioned techniques at the earlier stage, existing analytical preference models also include value-based models (Cook & DeVor, 1991), agent-based models (Zhang et al., 2011), and network-based models (M. Wang, Huang, et al., 2016; M. Wang, Chen, Fu, et al., 2015; M. Wang, Chen, Huang, et al., 2016). Among value-based models, random utility theory has been the predominant approach for modeling customer preferences (H. Q. Chen et al., 2013). This theory suggests that a customer's choice is determined by comparing the utilities of different alternatives, which depend on both the product attributes of competing design alternatives and the customer's individual characteristics. Specifically, **Discrete Choice Analysis** (DCA) (K. E. Train, 2009) and conjoint analysis (Tovares et al., 2013) have been extensively adopted by the design research community (Frischknecht et al., 2010; L. He et al., 2014; Hoyle et al., 2010).

**Discrete Choice Analysis (DCA)** has its origins in economics but has subsequently been extended to fields such as transportation research (M. E. Ben-Akiva & Lerman, 1985; Sha et al., 2016), engineering design (Sha, Wang, et al., 2017), systems engineering (H. Q. Chen et al., 2013; Sha & Panchal, 2014), and numerous other disciplines to address the need for estimating individual preferences and general system (market) demand. It should be noted that statistical analysis (Box & Tiao, 2011; Green et al., 1976; R. A. Johnson & Wichern, 2002; Neter et al., 1996) and data mining/machine learning techniques (Bishop, 2006; Witten & Frank, 2002) also have a long history of use in market research (Allenby & Rossi, 1998; Berry, 2004; Lilien et al., 1995) and engineering design (H. Q. Chen et al., 2012; Malak & Paredis, 2010; Ren & Papalambros, 2011; Tucker & Kim, 2008, 2009, 2011; L. Wang et al., 2011) for analyzing customer preferences. However, most of those techniques are aggregate approaches, meaning that they are more suitable for modeling group preferences when considering similarities among customers rather than individual preferences. Consequently, there is a need for employing Discrete Choice Analysis to address individual preferences more effectively.

The fundamental part of approaches applying random utility theory is formulating the utility function. In Discrete Choice Analysis (DCA), a decision maker's utility of selecting alternative $i$, denoted as $U_i$, is composed of two parts: the observed utility $V_i$, which is deterministic and fixed from the researcher's perspective, and the unobserved utility $\epsilon_i$, accounting for uncertainties like unobserved variations, measurement errors, and function misspecifications. This relationship is expressed as:

$$U_i = V_i + \epsilon_i \tag{2.1}$$

In a DCA, $V$ is modeled as a function of explanatory variables, in a linear additive form (M. E. Ben-Akiva & Lerman, 1985), as shown in Equation 2.2.

$$V_i = \boldsymbol{x_i} \boldsymbol{\beta_i^T} = \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \cdot + \beta_{in}x_{in} \tag{2.2}$$

where $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \cdot, x_{in})$ is a vector of $n$ variables and $\boldsymbol{\beta_i} = (\beta_{i1}, \beta_{i2}, \cdot, \beta_{in})$ is the vector of model parameters that quantify preferences in decision-making. DCA is derived based on random-utility maximization, meaning that the alternative $i$ is chosen over $j$ if and only if $U_i \geq U_j, \forall i \neq j$. Thus, the choice probability of alternative $i$ is:

$$P_i = P(U_i \geq U_j) = P(V_i - V_J \geq \epsilon_j - \epsilon_i) \quad \forall i \neq j \tag{2.3}$$

Since $P_i$ is the cumulative distribution of $\epsilon_j - \epsilon_i$, Equation 2.3 can be solved once the density function$f(\epsilon)$ is specified because. Different DCA models can be generated based on the choice of $f(\epsilon)$. For example, if $\epsilon$ is assumed to follow the Gaussian distribution, the resulting DCA model is known as the probit model. On the other hand, if $\epsilon$ is assumed to be identically and independently distributed following the Gumbel distribution, the resulting DCA model is known as the logit model (M. E. Ben-Akiva & Lerman, 1985). Furthermore, in scenarios where one chooses an alternative from multiple options, the logit model becomes a multinomial logit model. The choice probability of alternative $i$ is:

$$P_i = \frac{e^{x_i \beta_i^T}}{\sum_{j=1}^{J} e^{x_j \beta_j^T}} \tag{2.4}$$

In addition to the multinomial logit model, there are other DCA models such as nested logit (Kumar, Chen, et al., 2009) and mixed logit (Hoyle et al., 2010) that have been developed. These models can capture the system heterogeneity and the random heterogeneity among individuals by introducing customer attributes and using random coefficients respectively. Despite the utility-based approach's solid foundation in modeling customer preferences, it has several drawbacks, including dependency, rationality, and the necessity for choice sets. Dependency refers to the fact that standard logit models in DCA assume independence by ignoring correlations in unobserved factors across product alternatives. This implies that a customer's choice of one product is not influenced by adding or substituting another product in the choice set, which is often an unrealistic assumption. Rationality concerns the presupposition that utility functions are based on customers making rational decisions. However, in reality, their choices can be influenced by others and may appear ”irrational”. Lastly, the necessity for choice sets highlights that in cases where choice set data is lacking, misspecification of choice sets can lead to less accurate choice model estimates. This is particularly true for products with a wide range of alternatives (Shocker et al., 1991; Williams & Ortúzar, 1982).

To overcome these limitations, some studies have started investigating the use of statistical network models for estimating customer preferences (J. Fu et al., 2017; Sha, Huang, Fu, et al., 2018). Among the available network-based modeling techniques, the Exponential Random Graph Model (ERGM) has emerged as a particularly promising approach (Snijders et al., 2006). ERGM provides a versatile statistical inference framework that models the influences of both exogenous effects (e.g., nodal attributes) and endogenous effects (network structures/nodal relations) on the likelihood of connections between nodes. Applications of ERGM include studying customers' consideration patterns (Sha, Wang, et al., 2017), evaluating the effects of technological changes on market dynamics (M. Wang, Huang, et al., 2016), examining customers' sequential consideration and choice behaviors (J. Fu et al., 2017), and forecasting products' co-consideration relationships (Sha, Huang, Fu, et al., 2018; M. Wang et al., 2018). Further details on network-based modeling of customer preferences will be discussed in section 2.2 and section 2.3.

### 2.1.3 Multi-stage models: consideration-then-choice

Considering the various methods that have been investigated for customer preference modeling, it remains essential to introduce the two-stage decision-making process, a critical concept in effectively modeling customer preferences. This process, which is composed of the consideration stage and the choice stage, is widely recognized in customer research and marketing literature (Hauser & Wernerfelt, 1990; Roberts & Lattin, 1991; Shocker et al., 1991). At the consideration stage, a customer forms a consideration set from the myriad of available market options (Hauser & Wernerfelt, 1990), and then at the choice stage, the customer makes evaluations and finally selects a product from their consideration set based on individual preference and product attributes (Roberts & Lattin, 1991). The significance of introducing the two-stage decision-making process lies in its ability to better capture the nuances of customer behavior, enabling a more accurate understanding of the factors that influence their preferences during both the consideration and choice stages.

Comprehending the dynamics of these stages is crucial for devising effective marketing strategies and forecasting customer behavior.

Numerous studies have been conducted to investigate the determinants of consideration and choice stages in the decision-making process. Hauser and Wernerfelt (1990) explores the concept of consideration sets and proposes a model for evaluating the cost of including or excluding certain products from a consideration set. Roberts and Lattin (1991) developed a model of consideration set composition that explores how consumers form their consideration sets when making purchasing decisions, which sheds light on the factors that influence which brands are included in consumers' consideration sets and how many brands are considered. Chatterjee and Eliashberg (1990) integrated social network influence in their analysis, highlighting the role of interpersonal communication and peer effects in shaping consumer preferences during both the consideration stage and the choice stage.

Despite the existing research, including screening methods and heuristics from MacDonald et al. (2009) and Shin and Ferguson (2017), there is a need for quantitative approaches to understand the factors that drive customers' consideration decisions and investigate the difference between such factors and those that drive their purchase decisions. Additionally, most studies have primarily focused on final choice decisions, presuming fixed consideration sets for each customer, which do not reflect most realistic scenarios (Hauser & Wernerfelt, 1990). Consequently, understanding customers' consideration preferences is as vital as comprehending choice preferences. Further investigation and literature support will bolster the development of models that accurately capture the nuances of the two-stage decision-making process.

## 2.2 Data-driven network analysis

Having emphasized the importance of customer preference modeling in engineering design, this section discusses data-driven network analysis applied in product design and market study. We first present the applications of network analysis in the engineering design field and its significance. Next, we introduce different types of networks and their respective applications. Finally, we examine prevalent network metrics and descriptive analyses, which serve as fundamental instruments for quantifying and interpreting complex network structures. This discussion paves the way for the subsequent section on analytical methods in network modeling and their applications in product design research.

### 2.2.1 Importance and applications of network analysis in engineering design

In Section 2.1, we discussed the numerous challenges encountered when modeling customer preferences using traditional methods. One primary challenge is the considerable uncertainties associated with customers' decision-making processes, which are influenced by factors such as market demand, societal norms, and technological innovation. Additionally, the proliferation of social media has introduced new forms of social interactions, such as online reviews, which further exacerbate these uncertainties (Brock & Durlauf, 2001). The complex nature of the decision-making process itself also poses challenges, particularly in the context of the multi-stage (consideration-then-choice) framework discussed in Section 2.1. Furthermore, difficulties emerge when modeling heterogeneous human behaviors, intricate human interactions, and extensive product options. Network analysis addresses these challenges by capturing the interdependencies among entities, specifically customers and products.

Network analysis has emerged as a crucial method for statistical analysis of engineering systems across a broad spectrum of scientific, social, and engineering fields (Albert et al., 2000; Barabási & Albert, 1999; Braha et al., 2006; Holling, 2001; Hoyle et al., 2010; M. E. Newman, 2003; Simon, 1977; Wasserman & Faust, 1994). The premise underlying the network-based approach is that customer-product relationships, akin to other engineering systems that exhibit dynamic, uncertain, and evolving behaviors, can be considered as complex socio-technical systems, which are analyzed using social network theories and methods (W. Chen et al., 2020). By investigating the structural and topological features in customer-product networks, it is possible to understand the patterns of customer-product interactions while taking into account the heterogeneity among customers and among products.

In their recent study, Sha et al. (2019) conducted a comparison between network-based and utility-based approaches for customer preference modeling. They found that a network-based statistical model provides consistent results and identical factor effects as the discrete choice analysis method when only exogenous variables are considered. This finding emphasizes the advantages of statistical methods for network modeling as a comprehensive framework. These methods not only encompass the utility-based approach but also account for complex endogenous relationships within the design ecosystem, which are inadequately addressed by the utility-based approach alone. Moreover, network-based methods come with other analytical tools. For example, network graphs are used to visualize complex relationships (links) between individuals (nodes). In addition, descriptive analysis is used to quantify customer preferences or product markets.

### 2.2.2 Networks with different levels of complexity

In the following section, we introduce the different types of networks, with varying levels of complexity, that have been explored in the context of customer-product relations. These networks can be classified into three primary categories: *unidimensional* network, *bipartite* network, and *multidimensional* network. Figure 2.1 provides a visual representation of these network structures. Among these structures, both bipartite and multidimensional networks model customers and products as separate nodes, with the relationship between customers and products (customers consider and choose products) represented as links. In contrast, unidimensional networks concentrate on product competition based on aggregated customer preferences. In unidimensional networks, nodes represent products in the market, and links are formed based on customers' co-consideration or choices between products.

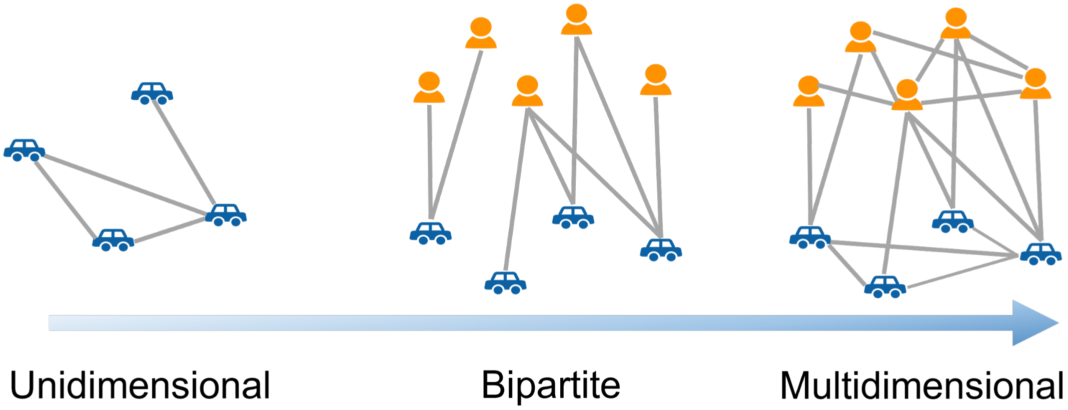


Figure 2.1: Unidimensional, bipartite, and multidimensional networks in customer-product relation modeling.

Prior research has demonstrated the importance of modeling unidimensional networks for customer preference analysis in the context of product design and market study. For instance, Sha, Wang, et al. (2017) investigated a binary unidimensional network to comprehend the influence of endogenous effects, such as existing competitive relationships between car models, on the formation of new competitive relationships in the market. Based on co-consideration relations that reflect customers' aggregated preferences, Xie et al. (2020) further investigated the dynamic evolution of the unidimensional network. A *bipartite* network, which specifies the relationship types (consideration or choices) between the customer and the product layer, but not within each layer, effectively models customers' decision-making processes. This approach allows for a more nuanced investigation of customer-product interactions. In their work, J. Fu et al. (2017) conducted a two-stage customer preference modeling study that demonstrated how product attributes could have different effects on customers' considerations and choices within the car market. Bi et al. (2021) also employed bipartite network methods, focusing on multi-year data analysis. To further capture the influence of customer social networks and product association on market competition, M. Wang, Chen, Huang, et al. (2015) introduced a *multidimensional* customer-product network. In multidimensional networks, links within each layer and between the layers are considered simultaneously. This approach used a more comprehensive representation of customer-product relationships in the context of customer preference analysis. A summary of descriptions and representative literature of various customer-product network levels can be found in Table 2.1.

| **Network Type** | **Description** | **Reference Paper** |
|---|---|---|
| Unidimensional | A unidimensional network with only product nodes can describe the product competition relationship based on aggregated customer preferences. | Sha, Huang, Fu, et al., 2018; Sha, Wang, et al., 2017; M. Wang, Huang, et al., 2016; M. Wang et al., 2018; Xie et al., 2020 |
| Bipartite | A bipartite network defines the relation (consideration or choices) between the customer layer and the product layer but not within each individual layer. | Bi et al., 2018; Bi et al., 2021; X. Fu et al., 2017 |
| Multidimensional | A multidimensional customer–product network, where the links within each layer and between both layers are considered in one network, captures the influence of customer social networks and product association (e.g., product family) on market competition. | M. Wang, Chen, Huang, et al., 2015, 2016 |

Table 2.1: Comparison of different network structures in customer-product relations

### 2.2.3 Network notions and descriptive analysis

To effectively conduct network analysis, it is crucial to become familiar with the associated terminology and metrics. This section introduces the common terms and metrics used in network analysis, and shows how they are applied in the context of customer-product networks.

A network can be represented as a **mathematical graph** $G(N, E)$, consisting of a **node set** $N = 1, 2, \cdots, n$ and an **edge (link) set** $E$ that includes all links between node pairs. Links can be either undirected, denoted by an unordered pair $i, j$, or directed, represented by an ordered pair with sender node $i$ and receiver node $j$. An $N \times N$ **adjacency matrix** provides a mathematical representation of a graph, where cells indicate the presence (1) or absence (0) of a link. Adjacency matrices for undirected and directed networks are symmetric and asymmetric, respectively. The adjacency matrix described above applies to unidimensional networks and can be easily extended to bipartite networks. The adjacency matrix of a bipartite network can be represented by an $N \times M$ matrix, where $N$ and $M$ denote the number of nodes of each type, respectively. Non-zero elements in the matrix indicate links between corresponding nodes. Furthermore, the links of the network can be weighted, which corresponds to a valued adjacency matrix in which each cell's value signifies the strength of a relationship.

In addition to the fundamental notion of a network, several key concepts are instrumental in characterizing and analyzing networks. These concepts encompass network density, degree distribution, connectivity metrics, and centrality indices. Moreover, network communities, which represent closely connected clusters of nodes, play a crucial role in understanding social networks. Consequently, community detection algorithms have been developed to identify these structures effectively. A summary of important network concepts and their implication in customer product networks can be found in Table 2.2.

Descriptive network analysis can be applied at various network levels to offer insights into com-

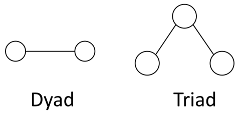
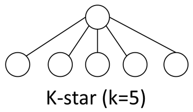
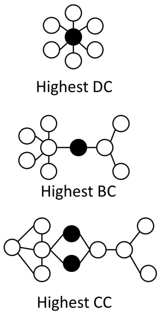
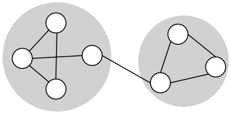
| Concept | Description | Mathmatical or Graphical Representation | Implication in Customer-Product Networks |
|---|---|---|---|
| Network Density | Ratio of the number of links to the number of possible links. | $D_{undirected} = \frac{2L}{N(N-1)}$, $D_{directed} = \frac{L}{N(N-1)}$ ($L$ is the number of links, $N$ is the number of nodes) | Identifies whether the competition or customer preference network is sparse or dense |
| Dyad and Triad | Dyad: Pair of nodes and the possible tie between them.<br>Triad: Three-node structure and the possible connections among them. |  | Analyze customer-product interactions and product competitions. |
| Geodesic Distance | Shortest path length between two nodes. | $d_{ij}$ | Determines the proximity between customers and products. |
| Degree Distribution | Distribution of node degrees across all nodes in the network. | $P(k) = \frac{N_k}{N}$ ($N_k$ is the number of nodes with degree $K$) | Reveals popular products and active customers. |
| K-star | Network subgraph centered on a node with k links from the node to k other nodes. |  | Reveals popular products and active customers. |
| Centrality Indices | Degree centrality (DC) quantifies a node's importance based on its link count, betweenness centrality (BC) assesses a node's presence in shortest paths between other nodes, and closeness centrality (CC) evaluates a node's proximity to all nodes within the network. |  | Identifies influential customers and popular products. |
| Clustering Coefficients | captures the degree to which the neighbors of a given node link to each other | Local clustering coefficient: $C_i = \frac{2T_i}{k_i(k_i-1)}$ ($k_i$ is the degree for node $i$, $T_i$ is the number of triangles on node $i$) | Quantifies the interconnectedness of the competitive relationships among the products that are related to the focal product. |
| Community Detection | Grouping nodes to form densely connected internal structures and sparse connections with other node sets. |  | Discovers product clusters and customer segments. |

Table 2.2: Summary of network concepts and their implication in customer-product networks

plex systems. In a unidimensional network, it is used to study product competition. For example, M. Wang, Huang, et al. (2016) created a unidimensional network of cars and discovered that the Audi FAW Q5 and Ford Kuga are popular vehicles, which have high ranks in degree centrality and in-degree hierarchy. Conversely, the Volvo V40 and Ford Edge are frequently considered during car purchases (high-degree centrality in an undirected network) but lag in customers' final choices (low in-degree hierarchy). In multidimensional networks, descriptive network analysis examines the interactions between different system levels (customers and products), such as peer influence — where customers who are linked to each other tend to choose the same products. By understanding these complex interactions, businesses can identify potential opportunities for improving their products and developing strategic marketing campaigns that enhance a product's reputation and appeal to customers. Therefore, descriptive network analysis serves as a valuable tool for designers, marketers, and decision-makers to determine product positioning, prioritize product features, and inform strategic planning at various stages of a project.

## 2.3 Analytical methods in network analysis

In this section, we aim to introduce the fundamental analytical tools and techniques commonly employed in network analysis, with a focus on two prominent modeling techniques: Exponential Random Graph Models (ERGM) and Graph Neural Networks (GNN).

### 2.3.1 Statistical network models: Exponential random graph models (ERGM)

Moving beyond traditional descriptive network analysis, statistical models like exponential random graph models (ERGM) can serve as a comprehensive and integrative statistical inference framework for interpreting complex preference decisions. ERGMs have been utilized to study customers' consideration behaviors in unidimensional networks at the aggregated market level

(Sha, Wang, et al., 2017) and in multidimensional networks at the disaggregated customer level (M. Wang, Chen, Huang, et al., 2016), respectively. The estimated unidimensional model has been applied to forecast the impact of technological changes (e.g., turbo engines) on market competition (M. Wang, Huang, et al., 2016), demonstrating the advantages of employing a network-based preference model for design.

To employ an ERGM for statistical inference, we define adjacency matrix $Y$ as a random variable to represent the graph. Then we use $Y_{ij}$ to denote the existence (1) or absence (0) of a tie between nodes $i$ and $j$ or from $i$ to $j$, and $y_{ij}$ for an instantiation of $Y_{ij}$. The full network may be denoted with instantiation $\boldsymbol{y}$. Network nodes (actors) can possess attributes of various types, representing individual-level properties such as product specifications and customer socio-demographics. Nodal attributes can be binary, categorical, or continuous. The ERGM models the probability of observing the graph $\boldsymbol{Y}$ as follows:

$$Pr(\boldsymbol{Y} = \boldsymbol{y}) = \frac{exp(\boldsymbol{\theta^T g(y)})}{\boldsymbol{\kappa}(\boldsymbol{\theta})} \tag{2.5}$$

In Equation 2.5, $\boldsymbol{y}$ represents the observed network, a random realization of $\boldsymbol{Y}$; $\boldsymbol{g(y)}$ is a vector of network statistics corresponding to network structural characteristics in $y$, node attributes, and edge attributes; $\boldsymbol{\theta}$ is a parameter vector indicating the effects of the network statistics; $\boldsymbol{\kappa}(\boldsymbol{\theta})$ is the normalizing constant that ensures the equation is a proper probability distribution. Equation 2.5 suggests that the probability of observing any particular graph (e.g., MCPN) is proportional to the exponent of a weighted combination of network characteristics: one statistic $g(y)$ is more likely to occur if the corresponding $\theta$ is positive.

In terms of network statistic $g(y)$, ERGMs can estimate the effects of various network configurations to explain the observed relational data within social networks (Lusher et al., 2013). A few configuration examples are provided in Table 2.3. The network configurations can be grouped into two main categories: **exogenous effects** and **endogenous effects**. The objective of network analysis is often to interpret the meaning of these configurations to understand customer-product relationships for engineering design. **Exogenous effects** posit that the attributes of products or customers can influence potential tie formations in a given structure. At a dyadic or two-node level, interpretation resembles the attribute effect in a logistic regression (Strauss & Ikeda, 1990; Wasserman & Pattison, 1996). Main effects can be used to assess the attractiveness of a product attribute, while interaction effects capture whether specific features are favored by particular customer groups. Product association relations can be captured by homophily effects that integrate customer preferences with product similarities, enabling the analysis to explain whether certain customer types tend to consider product alternatives associated with specific attribute sets. **Endogenous effects**, which are pure structural configurations, pertain to well-known structural regularities in the network literature. For example, the edge effect represents network density, the star effect highlights central nodes, the cross-level association-based closure effect captures structural patterns in relation to attribute homophily, and the cross-level ”peer influence” effect examines how connected nodes influence each other’s preferences.

ERGMs provide a flexible approach for modeling complex network structures, accounting for the interdependence of network links rather than assuming independence. This framework is capable of incorporating various nodal attributes, including binary, categorical, and continuous, to ascertain their association with the formation of network links. Moreover, ERGMs can effectively characterize both local and global network features, providing some degree of flexibility when working on diverse types of networks and relational data. The input data of ERGMs can be either cross-sectional or longitudinal, allowing for the construction of dynamic models that capture the evolution of networks over time. In contrast to machine learning models that prioritize prediction tasks, ERGMs serve as explanatory models, enabling the quantification of social influence

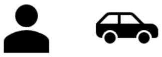
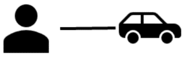
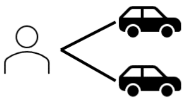
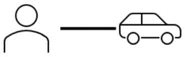
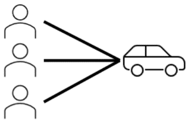
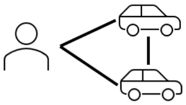
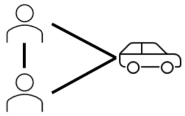
| Effects | Structures | Interpretation of Influences |
|---|---|---|
| **Exogenous effects** | | |
| Main attribute effect |  | Value of a customer or product attribute on link probability. Example: Fuel-efficient cars are more likely to be considered. |
| Interaction effect |  | The interaction between attributes of two types of nodes on link probability. Example: Customers from large families tend to consider larger-sized cars. |
| Homophily effect |  | Similarity of the attributes of two products on their probability of connecting to one customer. Example: Two cars with similar prices tend to be considered together by the same customer. |
| **Endogenous effects** | | |
| Baseline propensity (Edges) |  | Baseline probability of a customer considering or choosing a product at random. |
| Star effect (Alternating k-stars) |  | Impact of the-rich-get-richer. Example: The network links are centralized around a few high-degree (popular) product nodes. |
| Closure effect |  | Whether a closed structure is more likely to occur involving two product nodes with an association link. Example: Customers tend to consider two cars with many common features at the same time. |
| Peer influence |  | Whether a closed structure is more likely to occur involving two customers with a social influence link. Example: Customers tend to be influenced to choose the product that their ”peers” recommend. |

Table 2.3: ERGM network statistics for customer-product relationships: exploring exogenous and endogenous effects with cars as examples. Solid icons indicate customers and products with attributes, while hollow icons represent the absence of attributes.

and market structure effects by estimating the extent to which structural tendencies impact the likelihood of observing a specific network.

### 2.3.2 Deep learning-based network models (GNN)

Network data can be naturally represented by a graph structure that consists of nodes and links. Recently, graph neural networks (GNNs) have gained popularity due to their ability to model both discrete and continuous representations, showcasing a high level of expressive power. Consequently, they have been widely applied in domains that can leverage graph structures derived from the data (Battaglia et al., 2018). GNNs offer fundamental advantages over traditional unstructured machine learning methods, including enhanced interpretability, causality, and inductive generalization. The development of graph representations, reasoning, and prediction has led to remarkable progress in various applications, such as drug discovery (Jiang et al., 2021), image classification (Marino et al., 2016), natural language processing (L. Wu et al., 2023), and social network analysis (Fan et al., 2019). Notable examples of GNN implementations are related to recommendation systems (X. Wang et al., 2019; Ying et al., 2018), including Uber Eats (Jain et al., 2019), which employs GNNs for food item and restaurant recommendations, and Alibaba, which utilizes GNNs to model millions of nodes for product recommendations (J. Wang et al., 2018). Although the use of GNNs within engineering design is less prevalent, recent research has explored their application in product tolerance design (Li et al., 2021), machining feature recognition (W. Cao et al., 2020), and mechanical device functionality analysis (J. Wang et al., 2020). The success of these implementations has motivated us to investigate GNNs for studying customer-product relationships.

Graph-based machine learning tasks in networks address diverse challenges by capitalizing on the unique structure and properties of networks. These tasks encompass node classification, wherein labels are assigned to nodes based on their attributes and local neighborhoods (Kipf &

Welling, 2016); link prediction, which aims to discover potential connections between unlinked nodes (Liben-Nowell & Kleinberg, 2007); community detection, a technique to identify clusters of densely connected nodes sharing similar properties (Fortunato, 2010); network similarity, a measure used to compare the similarity or alignment between two networks (Conte et al., 2004); anomaly detection, a method to pinpoint unusual patterns or behaviors deviating from expected norms (Akoglu et al., 2015); and attribute prediction, a process to estimate missing or unknown attributes of nodes or edges (Grover & Leskovec, 2016). Collectively, these graph-based machine learning tasks offer a comprehensive framework for tackling complex issues in various domains, thus enhancing the efficiency and effectiveness of network analysis and prediction (Zhou et al., 2018).

**Graph representation learning** In a graph, each node is characterized by its features and the neighborhood of connected nodes. A node's behavior is often influenced by both its features and its nearby nodes, making the representation of nodes in graphs a challenging task. It is essential to learn meaningful graph representations that capture both local and global structural information as well as node feature information, considering the high-dimensional and non-linear nature of graph data. Graph representation learning methods address this challenge by enabling the automatic discovery of a node's vector representation, capturing both its graph structure and features from raw data. The result is a node embedding that can be interpreted as the node's learned features (or attributes). Ideally, similar nodes—those with comparable neighbors, connectivity, and features—should have analogous node embeddings. In a co-consideration network, two nodes can uniquely define an edge, allowing edge embeddings to be calculated using the corresponding node embeddings. By employing a suitably defined loss function in a machine learning model, it is possible to encourage all edges to exhibit edge embeddings more similar to non-existent (negative) edges. Therefore, learning the representation of nodes in a graph, known as node embedding, is a critical component for downstream tasks such as classification and regression.

There are two major classes of embedding algorithms: transductive learning and inductive learning. Transductive learning estimates the values of some nodes and edges while knowing the ground truth of the remaining nodes and edges in the graph. It involves predicting unknown nodes and edges by using supervised learning with known nodes and edges. Node embedding models, such as those employing spectral decomposition (Atwood & Towsley, 2015; Kipf & Welling, 2016) or matrix factorization methods (S. Cao et al., 2016; Qiu et al., 2018), are transductive. Inductive learning, on the other hand, trains a model on a graph and then makes predictions for nodes and links on an entirely new graph. Although the transductive approach does not efficiently generalize to unseen nodes in the same graph (e.g., for dynamically evolving graphs) and cannot generalize across different graphs, it has been the most prevalent in practice. On the other hand, the GraphSAGE method, proposed in 2017 (Hamilton et al., 2017), is an efficient inductive approach that leverages the attributes of adjacent nodes of the new node to generate its representation. GraphSAGE aggregates features from a sample of a node's local neighborhood. Consequently, training a GraphSAGE model on an example graph can generate node embeddings for previously unseen nodes as well, provided that they have the same set of attributes as the training data (i.e., no new attributes are introduced). GraphSAGE is particularly advantageous for graphs with numerous node attributes, which is often the case for customer-product networks.

**Interpretation of graph neural networks** In addition to using ML models for prediction, it is essential in engineering applications to understand their learning process and thus examine how different inputs affect the outcome. Interpretable ML methods present as an effective method that explains or presents model results in a way that humans can understand (Doshi-Velez & Kim, 2017; Molnar et al., 2020).

Identifying feature importance is a type of interpretable ML method that can help achieve this goal. It indicates the statistical contribution of each feature to the underlying model (Du et al., 2019). Among the techniques that estimate feature importance, model-agnostic methods (Ribeiro et al., 2016) present more flexibility and can work with any ML models, as they treat a model as a black box and do not inspect internal model parameters. Graph Neural Networks (GNNs) are considered black-box ML methods; therefore, when explaining the results of GNN models, we utilize model-agnostic interpretable methods.

In our work, permutation feature importance measurement, a model-agnostic approach, is employed to quantify the importance of features. Originally introduced by Breiman (Breiman, 2001a) for random forests and subsequently developed by Fisher, this approach involves the random permutation of a single feature while keeping other features unaltered. A pre-trained machine learning model then makes predictions. If a feature is crucial, prediction quality significantly deteriorates upon permutation. The feature's importance is quantified by the change in the prediction evaluation metric (Altmann et al., 2010). The core principle of this method is determining the significance of a specific feature to a trained ML model's performance by examining the changes in prediction accuracy after feature permutation (Altmann et al., 2010). Permutation-based feature importance has been employed in various fields, such as bioinformatics (Putin et al., 2016), engineering (Matin et al., 2018), and political science (Farinosi et al., 2018), providing valuable insights of feature importance into ML models.

# CHAPTER 3
# WEIGHTED NETWORK MODELING APPROACH TO PRODUCT COMPETITION ANALYSIS

## 3.1 Introduction

Unidimensional networks have successfully garnered significant attention in previous research for their effectiveness in modeling customer preferences. In such networks, nodes represent products, and links among them are formed based on whether customers have co-considered the products together. For example, Sha, Huang, Fu, et al. (2018) employed a binary unidimensional network to understand the influence of endogenous effects, such as the existing competition relations between car models, on the formation of new competitions in the market. Similarly, Ahmed et al. (2021) proposed a graph neural network approach to predict the binary unidimensional relationships between products. Two primary benefits of a unidimensional network approach are evident: First, it offers an aggregated representation of customer preferences and demand at a market level, which in turn, provides valuable decision-making support for businesses. Second, in a unidimensional network, customers' considerations and choices can be modeled jointly at the market level by the introduction of directed links. Therefore, it enables the prediction of market shares of different products beyond merely studying product competitions, thereby serving the design of market systems.

Despite earlier attempts at using network models and theories in understanding the driving factors in customers' consideration and choice behaviors, existing studies have several limitations. First, the networks are simplified as binary networks, meaning that the weights or the strength of links are neglected. However, link strength is an important aspect of understanding product competition as well as customer preferences. This is because to probe into the question of *how much* a competition relation between two products could be changed because of the change of designs or customer preferences, the link strength must be explicitly modeled. Second, the vast majority of previous research involving network models in car competition analysis fails to incorporate directed networks when modeling the ultimate product choice decision, focusing instead on the initial stage of choice-making—customers' consideration decisions. Addressing these shortcomings, this research, to our knowledge, is the first to apply weighted networks, along with consideration (undirected networks) and choice (directed networks), in the study of product competition and customer preferences.

The novel approach proposed in this research hinges on valued-ERGM models that allow links between nodes to possess weights and to be either directed or undirected. Despite the widespread applications of network modeling techniques in different research areas, the valued-ERGM technique (Krivitsky, 2012) has received little attention in engineering research. Our research aims at acclimating and transferring this statistical modeling knowledge into the engineering design field for further understanding product competition relations. In a unidimensional car competition network, we study both customers' consideration and choice behaviors by establishing two types of networks as illustrated in Figure 3.1 – an undirected network, in which links represent the co-consideration relationship and a directed network, in which a directed link between the two products co-considered indicates the customers' aggregated preferences towards the final choice decisions.

The **objectives of this research** are: a) to develop an approach based on valued-ERGM to model product competition, as exemplified by the study on both weighted undirected co-consideration network and weighted directed choice network; and b) to evaluate the performance of valued-

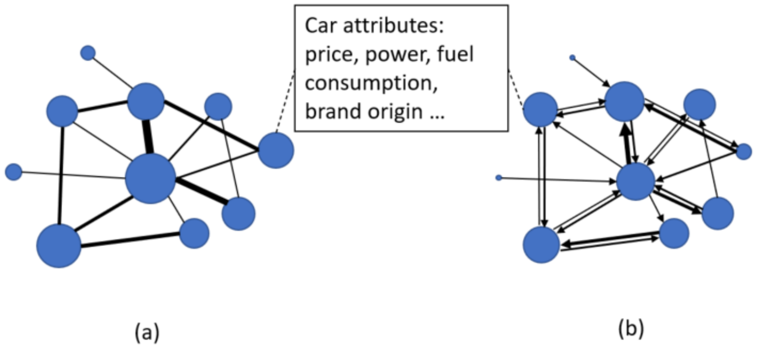


Figure 3.1: We use valued-ERGM network models to study product competition in both the consideration stage (the network in (a)) and the choice stage (the network in (b)). The nodes represent cars as an example in these network illustrations and links represent competition strength.

ERGM in link prediction (i.e., the competition strength prediction) when nodal attributes change in different years, e.g., the change of product design features when a car model upgrades from one year to another.

The primary contributions of this chapter are: first, a new network-based approach using valued-ERGM to explore product competition is proposed for the first time. Second, we demonstrate that valued-ERGM models predict customer consideration behavior substantially better than binary-ERGM models. Third, we show that valued-ERGM effectively models both directed and undirected networks in analyzing aggregated customer considerations and purchasing behaviors.

## 3.2 Weighted network construction and descriptive analysis

To capture the multi-stage nature of a customer's decision-making process, we build two different unidimensional networks, the "co-consideration network" and a "choice network". The first is an undirected network that represents customers' choice set in the consideration stage and the second is a directed network, which represents the customers' aggregated choice preferences.

In both networks, a product (in this case, a car) corresponds to a node. Each node is associated with a set of attributes like price, fuel consumption, and engine power. We denote both networks as $G = (V, \varepsilon, W)$, where $V$, $\varepsilon$ and $W$ represent nodes, links, and weights respectively. Figure 3.1 provides a simplified illustration for both the unidimensional consideration and the choice networks that we investigate. The thickness of the link between two nodes is proportional to its strength (i.e. the number of customers who co-consider the two products or choose one product over the other), and the size of the node is proportional to the popularity of the product (i.e. the number of customers who consider or purchase the product).

Our dataset contains survey data from 2013 and 2014 in the China market. In the survey, there were around 53,000 and 60,000 respondents respectively in 2013 and 2014, who specified which cars they purchased and which cars they considered, before making their final choice. Each customer indicated at least one, and up to three cars that they considered. The dataset also contains many car attributes, e.g., price, power, brand origin, and fuel consumption, and customer-specific attributes, e.g., gender, age, etc.

**Co-consideration network** To study car co-consideration, we start by creating a car co-consideration network based on customers' survey responses in the 2013 survey data. For purpose of validation, we control the studied market size and a random sampling of 50,000 customers was made. It is noteworthy that customers who have only considered one car in the survey are removed because they do not provide valuable information about product competition, and our network currently has taken roughly 38,000 customers. The network consists of 296 unique car models as network nodes. The link between a pair of nodes carries a weight equal to the number of customers who considered both car models together in their consideration set. The overview of the 2013 co-consideration network is shown in Figure 3.2. As the node size is proportional to the weighted degree of a car model, a larger node size depicts a more popular car model because it is considered by more customers.

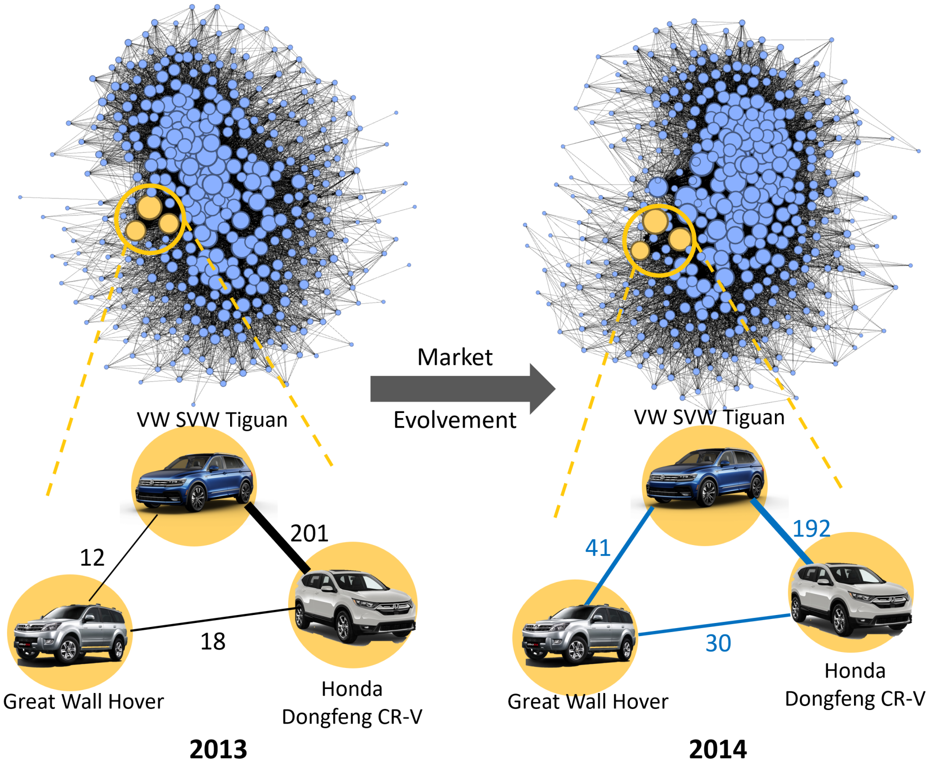


Figure 3.2: An overview of the 2013 and 2014 co-consideration network (Top): the blue nodes represent car models and black links represent co-consideration relations. The node size is proportional to the weighted degree of a car model and the link width is proportional to the link strength of the co-consideration relation. And an example of the local co-consideration network between three cars changing from 2013 to 2014 (Bottom).

Similarly, a thicker link width displays a stronger co-consideration relationship (competition) between a pair of cars. Figure 3.2 also shows a glimpse of a three-way competition. In this example, cars “Great Wall Hover” and “Honda Dongfeng CRV” appear together in the consideration set of 18 customers in 2013 and 30 customers in 2014, showing that their competition has potentially increased in one year (note the sampled market size for 2013 and 2014 are the same). In contrast, cars “VW SVW Tiguan” and “Honda Dongfeng CRV” appear together in the consideration set of 201 customers in 2013 and 192 customers in 2014. This shows that their competition has decreased in one year, although both car models are still more popular than the “Great Wall Hover”,

Table 3.1: Summary of 2013 Co-Consideration Network Descriptive Characteristics

| No. Nodes | Network Density | Ave. Strength | Ave. Degree | Ave. Weighted Degree | Global Clustering Coefficient |
|---|---|---|---|---|---|
| 296 | 0.152 | 5.323 | 22.355 | 118.80 | 0.616 |

as indicated by the sum of all link strengths connected to them.

Table 3.1 presents a summary of our network's descriptive characteristics. Network density, which calculates the portion of the potential connection between all nodes that are actually connected in a network, shows that among all possibly connected car models, $15.2\%$ of them are being co-considered, and an average of $5.323$ customers consider any connected car models indicated by the average strength. The average degree means that each car competes with $22.355$ cars on average. The average weighted degree indicates a car is co-considered with other cars by $118.80$ customers on average. The average global clustering coefficient of $0.616$ suggests that car models are very likely to engage in a multi-way competition.

**Choice network** In the case study of the choice network, we focus on the market competition among crossover SUVs, such as the Ford Edge and Mazda CX-7, which are designed with the body and space of an SUV but the platform of a sedan. This type of car model has gained increasing attention in recent years and has witnessed considerable growth in many countries, owing to the low cost, compact size, stylistic design, and better maneuverability. There are $14$ crossover SUV models in the 2013 survey data, and we have collected all survey data of which customers have either considered or chosen a crossover SUV model in that year. This gives a total of $1975$ customer observations. The directed choice network is established based on the customers' purchase behavior as described in the previous section and all competitors in the network are divided into four segmentation groups: Sedan, SUV, Luxury or Sport, and Crossover SUV. The visualization of the choice network is plotted in Figure 3.3, where the node size of a crossover SUV reflects the number of customers who have purchased it.

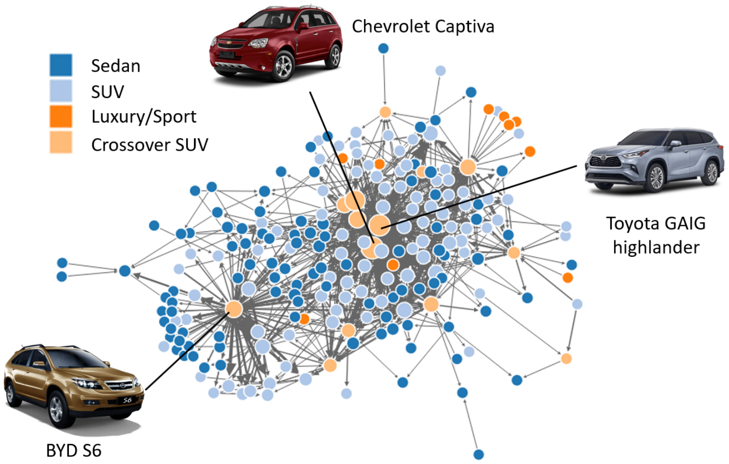


Figure 3.3: A force directed graph visualization of the 2013 choice network for Crossover SUVs. We observe that most crossover SUVs compete with Sedans and SUVs.

Overall, there are 217 car models in the crossover SUV choice network. All the links are directed and point to the "winner" in a competition. The average link strength is 2.431 corresponding to the average number of customers' purchases among all co-considered cars. A unique feature of the choice network is that the in-strength of a node is correlated with its market share. As illustrated in Table 3.2, and it lays the foundation of market share prediction using the choice network data.

### 3.3 Weighted network modeling and predictive analysis

With the established co-consideration and choice network (with link strength), we are able to use the extended version of the ERGM model, the valued ERGM model, to do the analysis.

Table 3.2: Market Share of Crossover SUV Segment

| Car Model | Node In-strength ($s_{in}(i)$) | Market Share ($\frac{s_{in}(i)}{\sum_{j=1}^{n} s_{in}(j)}$) |
|---|---|---|
| BYD S6 | 89 | 7.90% |
| Fiat Freemont | 88 | 7.82% |
| Ford Edge | 77 | 6.84% |
| Chevrolet Captiva | 258 | 22.91% |
| ... | | |
| The total | 1126 | 100% |

### 3.3.1 Valued-ERGM model

A limitation of traditional binary ERGM is that it cannot model networks with weighted links (e.g. the demand between two airports in an air transportation network). If one wishes to model a weighted network with the traditional ERGM, they have to first binarize the network with a link weight threshold. This process converts each edge to a binary $0$ or $1$ link so that the ERGM can take the resultant network as input. Researchers often use an artificial cut-off value and all the links with weights below the cut-off value are removed ($0$ links) and the remaining are kept ($1$ links). This dichotomization step may lead to biases and information loss, which can eventually affect network prediction.

Valued-ERGM (Krivitsky, 2012), a technique recently developed by statisticians, addresses this limitation by modeling the strength of links rather than merely their presence or absence. For a given set of discrete variables, a valued-ERGM is expressed as:

$$Pr(\boldsymbol{Y} = \boldsymbol{y}) = \frac{h(\boldsymbol{y})exp(\boldsymbol{\theta}^T g(\boldsymbol{y}))}{\kappa(\boldsymbol{\theta})}, \ \boldsymbol{y} \in \boldsymbol{Y} \tag{3.1}$$

where most of the parameters are the same as those in Eq. 2.5, and the normalizing factor $\kappa(\theta, \boldsymbol{y})$ can be expressed as $\sum_{y' \in \boldsymbol{Y}} h(\boldsymbol{y})exp(\boldsymbol{\theta}^T g(\boldsymbol{y}))$, to make the function output a feasible probability value. Two major distinctions between the valued ERGM and the regular ERGM are the support $\boldsymbol{Y}$ term and the reference distribution $h(\boldsymbol{y})$ term.

Different from binary ERGMs, the support of a valued-ERGM is over a set of weighted networks, which is often infinite or uncountable (Krivitsky & Butts, 2013). One cannot enumerate all possible weighted networks with real-valued link strengths. Thus in a weighted network case, we need to consider what the strengths of connections are and how they are distributed. This brings in the need of specifying a reference distribution, which determines the sample space and baseline distribution of link values. The sample space is a set of possible networks given the size and density of the observed network, which depends on the maximum value of the tie between any two nodes. A reference distribution simply answers the question of what the link distribution might look like in the absence of any ERGM terms. The ability to model valued links has greatly advanced network research as it enables researchers to conduct more nuanced examinations of network structures. Moreover, similar to traditional ERGMs, valued-ERGMs are capable of modeling networks with both undirected links and directed links. Recent development of valued-ERGM has extended to the continuous link strength (Krivitsky & Butts, 2013), which broaden the application to various problems. Despite these benefits, valued ERGMs are still very much an exploratory area within statistical network analysis (Scott, 2016) due to computational difficulties.

Valued-ERGMs have been employed in various applications ranging from policy studies (Scott, 2016), organizational communication (Pilny & Atouba, 2018) to disease transmissions (Silk et al., 2018) and global migration (Windzio, 2018). An important step of using valued-ERGM is to first define meaningful links and a way to measure the link strength. The definition of link strength often depends on the domain, and in the past, researchers have determined it based on factors ranging from the level of interaction between two nodes (Scott, 2016), the strength of friendship (Pilny & Atouba, 2018), or the total duration of human contact (Silk et al., 2018). These links, although valued, are typically discrete in a small range such as $\{0, 1, 2, 3\}$. Existing methods in the social science area cannot be directly used in our study to model the valued product competition networks because: a) the link strength in a product competition network could have a substantially large range. This infinite sample space increases the complexity of the task of prediction; and b) existing studies mainly concentrate on interpreting the models, whereas we focus on both interpretation and prediction. The prediction of the network involves network simulation based on the estimated parameters, and it can also serve as a validation of the fitted model. Despite their complexity, there are two motivations behind using valued-ERGM models in this work: 1) they can model the magnitude of competition strength between products, thereby supporting car manufacturers' strategic decisions on product positioning. As the valued-ERGM will establish the functional relations between the car design features and the competition strength, the resulting model will be able to predict future market competition based on the change of certain car features, such as a design upgrade or design modification. 2) With more information captured, the valued-ERGM model should demonstrate a better link prediction accuracy compared to traditional binary ERGMs.

### 3.3.2 ERGM estimation and interpretation

**Co-consideration network** In the implementation of the valued-ERGM model, we assign the selected car attributes to network nodes and the occurrence of co-considerations to the link strengths. Based on the sample space of link strength (non-negative, integer, and not bounded), the available reference distributions are the Poisson distribution and Geometric distribution. In an empirical setting, the Poisson distribution provides a converged and legitimate result, therefore, we have chosen the Poisson distribution as the reference distribution.

The input variables can be divided into three categories: the network configuration effects, the main effects (Sha, Huang, Fu, et al., 2018) and the homophily effects. The whole set of input variables can be found in Table 3.3. We use the statistical network analysis package “Statnet” in R programming, in which the valued-ERGM is integrated (Handcock et al., 2019). The second column of Table 3.3 (i.e, “Weighted”) shows the estimated coefficients from fitting the valued-ERGM models. The sum/intercept variable serves as a constant term in valued-ERGM and it estimates the likelihood of two cars’ co-consideration strength without any knowledge about the cars’ attributes. All the input variables, except the main effect of power and the homophily effect of the power difference, are statistically significant at the level of significance of 0.05. As all variables are normalized to a similar order of magnitude, the differences in the coefficients denote their relative importance in the model fit. Among the main effects, the coefficient of import effect is negative, but the coefficients of brand origin from different countries are positive. This implies that customers tend to consider domestically made cars with foreign brands, such as Ford Changan Focus, and Honda Dongfeng Civic. Variables like price, power, and fuel consumption are not as important as the other main effects.

We observe that the coefficients corresponding to the homophily effects are mostly positive and significant. This indicates that the homophily effects may play an important role in forming the competitive relations between two car models, which verifies our common beliefs. Among the homophily effects, market segment matching and brand origin matching are significant. This may reveal that car models within the same market segment and the same brand origin tend to be co-considered by customers. Furthermore, a statistically significant large negative coefficient of price difference shows customers prefer to consider cars in a similar price range. This observation aligns with our intuition, as a customer may consider cars within his/her budget range.

**Choice network** The procedure of network modeling of a choice network shares many similarities with that of a co-consideration network using the valued-ERGM approach. However, as the

Table 3.3: Estimated Coefficients of the 2013 Co-consideration Network for a weighted network and binary networks

| Input Variables | Weighted | Binary 1 | Binary 2 | Binary 3 |
|---|---|---|---|---|
| **Network configuration effect** | | | | |
| Sum/Intercept | - 9.54*** | -7.24*** | -8.80*** | -11.72*** |
| **Main effect (nodal attributes)** | | | | |
| Import | - 1.46*** | -1.00*** | -1.31*** | -1.59*** |
| Price (log2) | 0.27*** | 0.18*** | 0.21*** | 0.25*** |
| Power (log2) | 0.05 | 0.07 | 0.01 | 0.04 |
| Fuel consumption | - 0.03*** | -0.06*** | -0.02 | -0.01 |
| Brand origin (the US) | 1.42*** | 1.17*** | 1.43*** | 1.72*** |
| Brand origin (Europe) | 1.11*** | 0.82*** | 1.09*** | 1.31*** |
| Brand origin (Japan) | 0.45*** | 0.46*** | 0.58*** | 0.67*** |
| Brand origin (Korean) | 0.75*** | 0.69*** | 0.90*** | 1.07*** |
| **Homophily effect (dyadic attributes)** | | | | |
| Market segment matching | 1.16*** | 0.76*** | 0.93*** | 1.10*** |
| Brand origin matching | 0.87*** | 0.82*** | 0.97*** | 1.13*** |
| Price difference (log2) | - 1.90*** | -1.35*** | -1.83*** | -2.25*** |
| Power difference (log2) | -0.06. | 0.12 | 0.04 | -0.12 |
| Fuel consumption difference | - 0.30*** | -0.26*** | -0.38 *** | -0.41*** |

Note: ***p<0.001

choice behavior is not symmetric between pairs of nodes, the model terms are further specified for inward nodes or outward nodes. Specifically, the main effects in Table 3.4 refer to the nodal attributes of the inward nodes, hence we can learn the important attributes of the “winners” and find possible reasons behind the popularity of a car model. Besides, we have added two network structural effects, “cyclical weights” and ”transitive weights”, which measure the triadic closure and refer to the links from $i \rightarrow j$ that have two-path[1] from $j \rightarrow i$ and from $i \rightarrow j$, respectively (Figure 3.4). More precisely, in the product competition market, it accounts for a hierarchical three-way competition. The cyclical weights refer to the case when customers prefer car $k$ than car $j$ and prefer car $i$ than car $k$, while preferring car $j$ than car $i$. The transitive weights refer to the case when customers prefer car $k$ than car $j$ and prefer car $i$ than car $k$, while preferring car $i$ than car $j$.

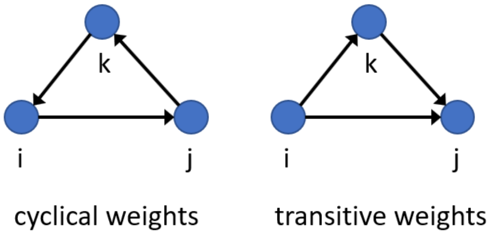


Figure 3.4: An illustration of cyclical weights and transitive weights. It refers to three-way competition in the market.

Table 3.4 shows the estimated coefficients from fitting three directed valued-ERGM models with different model terms. The first model is a baseline model with main effects and homophily effects, and the second and the third models include network structural effects to further investigate the endogenous network effect influence. Among all three networks, the estimated coefficients are consistent with small variations. In the choice network, the car models with lower prices, higher

[1] a two-path refers to a network structure that there are two edges connects from $i$ to $j$: $i \rightarrow h \rightarrow j$

Table 3.4: Estimated Coefficients of the 2013 Choice Network

| Input Variables | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| **Network configuration effect** | | | |
| Sum/Intercept | 6.19*** | 6.11*** | 5.77*** |
| Cyclical weights | | -0.06** | |
| Transitive weights | | | 0.16*** |
| **Main effect (inward node attributes)** | | | |
| Import | 0.03 | 0.03 | 0.03 |
| Price(log2) | -1.05*** | -1.04*** | -1.03*** |
| Power(log2) | 0.30* | 0.30* | 0.30* |
| Fuel consumption | 0.58*** | 0.58*** | 0.57*** |
| Brand origin (the US) | 0.82*** | 0.81*** | 0.78*** |
| Brand origin (Europe) | 0.15 | 0.13 | 0.15 |
| Brand origin(Japan) | 0.80*** | 0.80*** | 0.75*** |
| Brand origin(Korean) | 0.56*** | 0.56*** | 0.54 |
| **Homophily effect (dyadic attributes)** | | | |
| Market segment matching | 0.68** | 0.69*** | 0.67*** |
| Brand origin matching | 0.99*** | 1.00*** | 0.98 |
| Akaike Information Criterion (AIC) | -68209 | -68211 | -68252 |
| Bayesian Information Criterion (BIC) | -68113 | -68106 | -68147 |

[1] Note: .p <0.1; *p<0.05; **p<0.01; ***p<0.001

power, and higher fuel consumption are more likely to be bought by customers. This result is consistent with our common sense. Please note that for the group of customers who have a preference for crossover SUVs, they possibly prefer a model with higher fuel consumption which is usually in company with a higher power. Meanwhile, imported cars are not always preferred by this survey population, but a car with foreign brands still shows a positive effect on customers' final choice. Furthermore, the homophily effects have significant positive effects on the choice decisions, and the underlying reason is similar to the first case study. Also, in model 2 and model 3, the cyclical weights have a negative effect while the transitive weights have a positive effect. This implies that in a three-way competition, the competition relations tend to be transitive, meaning that if car A ”wins” a competition over car B, and car B ”wins” a competition over car C, then car A is likely to ”win” car C. Therefore, it can be inferred that the directed network market is hierarchical. We have also reported Akaike information criterion (AIC) and Bayesian information criterion (BIC) values for three models, a lower AIC and BIC value indicates a better model fit (Burnham & Anderson, 2004), and the models with network configuration statistics fit slightly better than the baseline model, which indicates that those network configurations could play an important role in the competition network formation.

### 3.3.3 Prediction on the future market

While statistical network models are typically used to interpret what factors lead to link formation or dissolution, predicting what a network will look like in the future is useful for manufacturers to make strategic decisions. In practice, if manufacturers can predict how the competition between car models would change when certain product design attributes are changed, they can use this knowledge to position their products in the market strategically against competitors. Using the estimated parameters of input variables in valued-ERGM, we can predict competition networks in the future, with new car attributes as input.

Based on the valued-ERGM equation Eq. 3.1, the distribution of network models is determined by a base network structure, estimated parameters, input variables, and a reference distribution. Therefore, when predicting a future competition network, we substitute the old car attributes with new ones and derive the distribution of the predicted network structures based on the valued ERGM formula. Then, we draw many samples from the network distribution (simulated networks) and take the averaged network structure as the aggregated network, which represents the central tendency (highest probable network) of all simulated networks. We use this aggregated network as our prediction and compare it with the known network in the future to show our model's accuracy.

Future predictions using aggregated simulations can be made for either the co-consideration network or the choice network. In the predicted co-consideration networks, the number of competitors and their strengths are predicted. In the predicted choice networks, the manufacturers will get an understanding of which car models are their main competitors.

**Co-consideration network** We start the model validation by performing simulations with the current network configurations and the estimated coefficients of the selected model terms. More concretely, we create $100$ simulated networks with the 2013 car co-consideration network configurations and the estimated parameters in Table 3.3, and then take the average of the link strength values from $100$ simulations and denote it as the aggregated simulated car co-consideration strength. The comparison of the link strength between the simulated network and the original network reveals the goodness of the model fit. Figure 3.5 (Top) plots the link strengths of the true network compared to the aggregated simulated network along the diagonal. We observe that two sets of link strengths are positively correlated, where a perfect $y = x$ line indicates a perfect fit. This is manifested by the Pearson coefficient of $0.988$ and the coefficient of determination ($R^2$) of $0.976$.

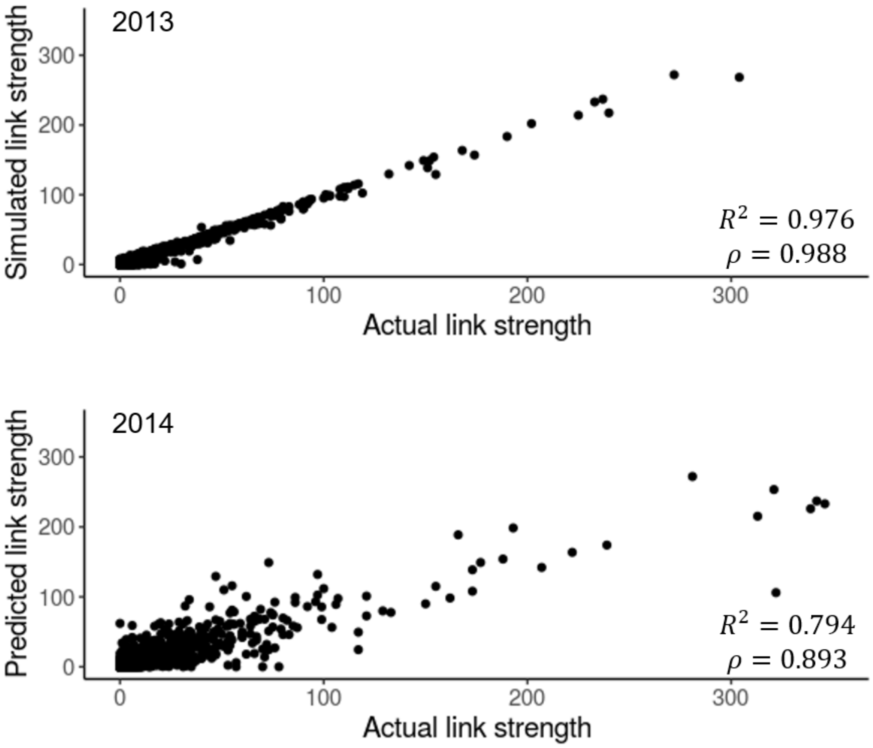
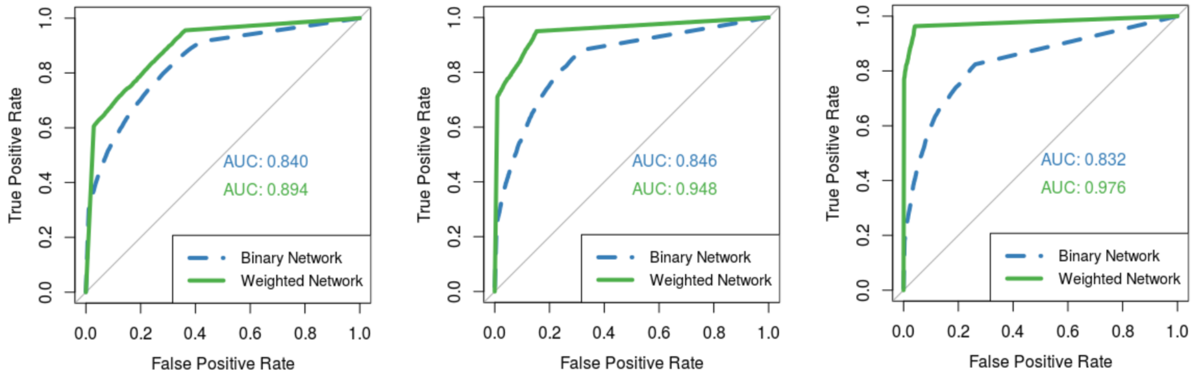


Figure 3.5: The goodness of fit using link strength comparison between the trained network and simulated network. Top: Link strengths of the trained network with the aggregated simulated network for 2013. Bottom: Link strengths of the true network with the aggregated simulated network for 2014 (unseen future data).

In practice, the benefit of training a statistical model is to predict the future state and behavior of networks that are unseen. While the market competition between different car models varies yearly, we test whether our fitted co-consideration model can be utilized to predict the co-consideration relationship in the future market. Figure 3.2 illustrates an example of the real market evolution. It can be observed that in 2014, Great Wall Hover gains more customers' consideration, and the strong co-consideration relationship between VW Tiguan and Honda CR-V decreases slightly. Our examination of the model's predictive power uses a similar method of network aggregation as used in the above validation study, but with the input of 2014 car attributes as the updated node attributes.

With a similar simulation process, we derive the aggregated predicted co-consideration network for the 2014 market data and compare it with the actual co-consideration network. The scatter plot of the actual link strength and the predicted link strength is reported in Figure 3.5 (bottom), with a $R^2$ of $0.794$ and Pearson coefficient of $0.893$. More importantly, we observe that although there exist some deviations between the prediction and the true link strength in the lower range of the link strength values, the prediction is better when the link strength is larger. In practice, the ability to correctly predict large link strength values is more important because they indicate more intense competition where major players in the market are always involved.

We want to further compare the prediction results with the previous binary non-weighted network baseline. However, for comparison, we have to convert a simulated weighted network to a binary counterpart using a cut-off value of the link strength.



(a) Cut-off = 1.0 (b) Cut-off = 2.0 (c) Cut-off = 4.0

Figure 3.6: Receiver Operating Characteristics (ROC) curves comparing the valued-ERGM model with binary-ERGM models with different cut-off values for network binarization on the 2014 car competition network. We observe that irrespective of what cut-off value is used, valued-ERGM models have higher precision and recall values than other models.

We choose three different cut-off values, $1.0$, $2.0$ and $4.0$, for creating the binary network. These cutoffs are determined based on the first, second, and third quantiles from the actual network link strength distribution. After that, we compare the predicted co-consideration network with the actual binary network. This comparison allows us to measure the false positive rate and true positive rate as metrics to evaluate the model performance. More specifically, we draw the Receiver Operating Characteristics (ROC) curve for each cutoff value. ROC curve (Fawcett, 2006) is a performance measurement for classification problems at various threshold settings of the predicted probability, and the larger the area under the curve (AUC) is, the better is a model's predictability.

For all the ROC curves, AUC for the weighted network is larger, which indicates a better predictive performance of valued-ERGM compared to binary ERGM. As the cut-off value increases, the performance of binary ERGM keeps, while the performance of valued-ERGM becomes better and better. This is because as the binary network becomes sparser, only links with higher strength are preserved and valued-ERGM has better performance in predicting those links.

**Choice network** In a directed choice network, the in-strength of node $s_{in}(i)$ is related to its market share. Hence, we can further validate the choice network by comparing the simulated market share for each crossover SUV with its true market share. Specifically, the in-strength fraction $\frac{s_{in}(i)}{\sum_{j=1}^{n} s_{in}(j)}$ is calculated based on an observed choice network for the actual market share of the crossover SUVs. Then, the simulated market share is derived by averaging the in-strength of the nodes from 100 simulations. The comparison of actual market share, simulated market shares of three different models, and the uniform market share (which assumes all crossover SUVs have the same market share and serve as a baseline) is plotted in Figure 3.7. Even though there exists a discrepancy for some car models (e.g., Mazda CX-9 and GM USA Buick Enclave), most of the predictions of car models show a consistent trend with the actual market share. Compared to the baseline of uniform market share, all simulated market shares have a $R^2$ value above 0.7, which indicates that more than $70\%$ of the observed variation can be explained by the fitted choice network model. Among them, model 1 has a $R^2$ value equal to 0.77, model 2 has a $R^2$ value equal to 0.70, and model 3 has a $R^2$ value equal to 0.74. As a side note, the models adding more network attributes do not provide a better-simulated market share than the baseline model (Model 1), which could be raised by the sparsity and less influence of the network structure.

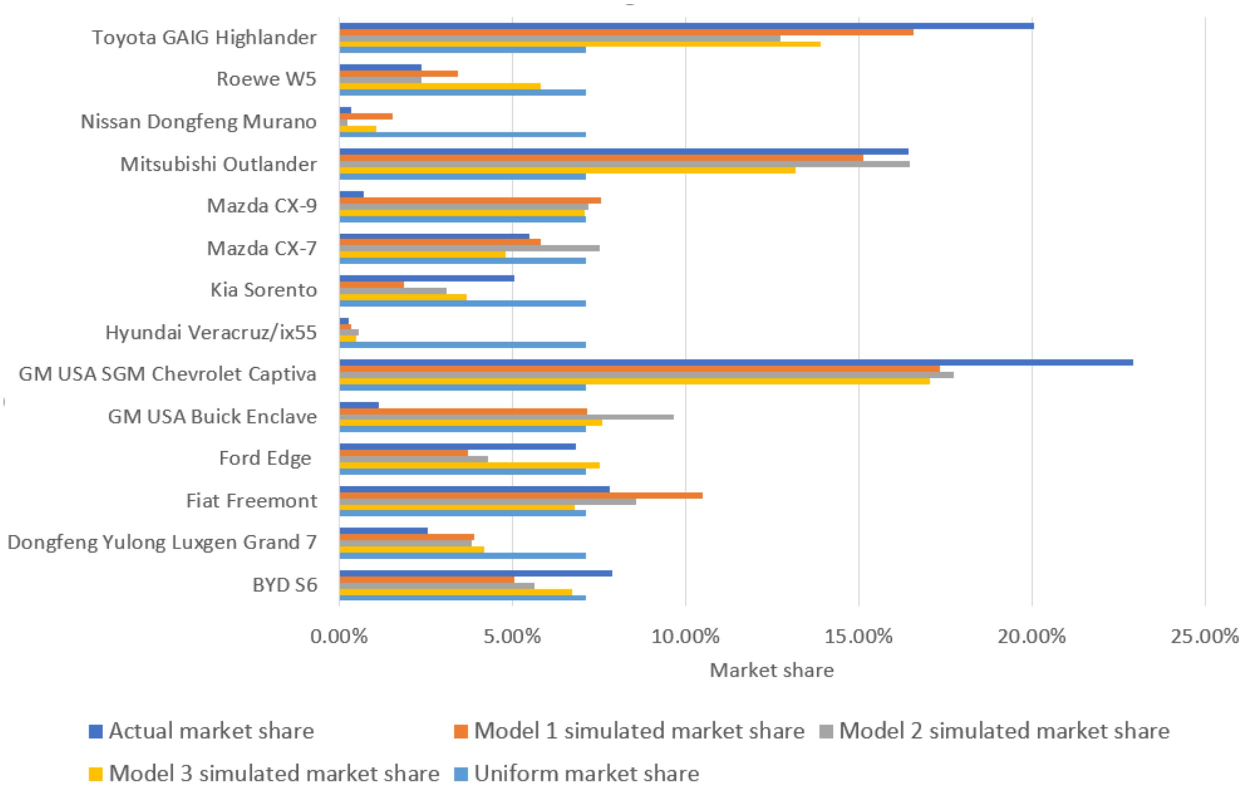


Figure 3.7: Valued-ERGM prediction of 2013 crossover SUVs market share aligns with the true market share.

While valued-ERGM shows a reasonably good fit for the relative pairwise competition and the market share, it does not predict well the absolute value of weights in the choice network. This is true in predicting both the current market and the future market. We suspect that this is due to the sparsity and directionality of the network. The network constructed in this case study only contains crossover SUVs, thus leading to a very low network density of 0.02.

## 3.4 Discussion

While the valued-ERGM model provides many advantages over existing statistical models, it is a relatively new model with a few theoretical and practical challenges that require attention and more research. In this section, we summarize the benefits and limitations of the valued-ERGM models and discuss how they pave the path to future research directions.

**Supporting engineering design decisions using valued-ERGM** One of the goals in using the valued-ERGM model is to demonstrate how the approach helps identify the important factors that influence product competition. These factors can support stakeholders in making strategic decisions. However, it is important to note that while the theoretical model allows one to estimate the importance of any attribute, the analysis in specific case studies may also depend on what product data is available and whether there indeed exists any relationship between product attributes and customers' choice decisions. To understand this, let us consider three hypothetical situations. In the first situation, a customer decides to buy a car merely based on the size of the car engine. Using a valued-ERGM model, the analysis results show that the size of the engine (or power – which is correlated with it) has a significant positive coefficient. In such a case, the network models inform that increasing the engine size can help gain a larger market share. However, increasing the engine size will inevitably increase the manufacturing cost, thus leading to a higher price. This, on contrary, may negatively influence the market share. There is obviously a trade-off decision the car manufacturer has to make, then the network model should help car manufacturers make decisions of choosing the right combination of design features.

In the second situation, we assume that a customer decides to buy a car merely based on the quality of its air-conditioning (AC) system. If the data we analyze does not include the AC design attribute, the results will not be able to provide specific insights into the impact of AC design on customers' choice behaviors. The only remedy for this is to collect data that captures the relevant attributes for the choice analysis. In the third situation, we assume that customers' choice behaviors are only influenced by social and/or cultural factors, but not car design features. In such cases, the coefficients of all design attributes may not have statistical significance. This indicates that improved design features may not help automakers gain more market share. Hence, the guidance provided to the manufacturer is to not waste resources on improving factors that do not have an impact.

In this chapter, the customers' choice behaviors described in the two case studies are a mixture of the three situations. For example, we find that some design attributes have a statistically significant influence, but we also discover that this dataset lacks information about certain car design attributes. Finally, many design attributes studied are not statistically significant, indicating that those attributes may not play a role in customer decisions.

From our current results for both case studies, we successfully identify a few factors that impact engineering design decisions for product consideration. Specifically, in the co-consideration network of case study 1 (Table 3.3), we observe that a car designer may want to reduce fuel consumption (which relates to engine efficiency) to increase the competitiveness of their car models. Although factors like price, power and fuel consumption are statistically significant, they do not directly provide actionable design guidance for a car manufacturer. In the choice network of case study 2 (Table 3.4), the model results help decision-makers with strategic planning. For example, in the crossover SUV market, the improvement of fuel consumption may not increase the likelihood of a car being purchased. Instead, reducing the price and increasing the power could be helpful to improve the market share. We notice that our dataset lacks certain car design attributes that may be influential to customers' choices. In future work, we aim to address this issue using crowdsourced data inputs. Moreover, we have discovered the important effect of particular network configuration statistics, such as “cyclical weights” and ”transitive weights” in Table 3.4 in the choice network. That manifests the advantages of (valued) ERGMs in utilizing network configurations to capture endogenous effects in a market. The insights into these endogenous effects help car manufacturers gain an in-depth understanding of the market and their competing opponents.

**Trade-off between feature engineering and model interpretability** In valued-ERGM models, we start with a large collection of features. These features can be node-specific (e.g., car fuel efficiency, price), link-specific (e.g., the price difference between two car models), or network-specific (e.g., popularity, density). The choice of what features to use has a large impact on the goodness of fit of the model, the estimated coefficients as well as their statistical significance. While we use automated methods for feature selection (which largely select features that are uncorrelated), the process is often manual. In contrast, one can use modern deep-learning models to learn hierarchical feature representations. Yet, the deep learning models are largely black-box and are hard to interpret, which is one of the key reasons for us to adopt the interpretable and theory-grounded statistical network models in this study. In the future, we will attempt to find the middle ground of reducing dependence on feature selection, while still retaining model interpretability by combining the two methods.

**Numerical issues with valued ERGMs** Existing literature reports two numerical issues of (valued) ERGMs: the reliability of model interpretation and computation issues for large networks.

**Reliability:** In recent years, there have been critiques of using (valued) ERGM packages related to the accuracy of inference methods reported by the statistical software for ERGM. While some experiments suggested that the variants of ERGM models can work well even with a relatively small sample taken from the network (A. D. Stivala et al., 2020), Shalizi and Rinaldo (Shalizi & Rinaldo, 2013) have argued that ERGMs are designed for modeling the entire network. In many applications, the data used consists of a sampled sub-network, which could lead to inconsistency of interpretation due to the MCMC sampling process. However, our first case study is unlikely to suffer from the reported issues due to two reasons: 1) the subset of customers in our network for the first case study only changes the link strength magnitude and we still use all nodes, 2) We also test with different subsets samples of customers and find that the results are similar, which indicates the reliability of our network models. For the second case study, we use a particular market segment of cars to create the network, which may suffer from reported limitations. Hence, we are cautious in generalizing our findings from the study on cross SUVs to other car segments.

**Computation issue for large/complex network:** It is reported in the literature (R. He & Zheng, 2013) that for large and complex network structures, the MCMC approach to estimate ERGM parameters may not converge. In our work, this limitation can be a problem for some stakeholders. There is some recent work on developing scalable binary ERGMs (An, 2016; A. Stivala et al., 2020), and the extension of such methods to valued-ERGMs can help alleviate the scalability problem for large datasets. Another approach that can improve the scalability of valued-ERGMs is to use kernelized approximate Bayesian computation. It can improve computational efficiency and is being adopted by popular packages (Yin & Butts, 2020) as an alternative to MCMC.

## 3.5 Conclusion

In this chapter, we enhance the network modeling approach for analyzing customer preferences and product competition by viewing customer-product relations in the context of a complex sociotechnical system. With a focus on the unidimensional network as the aggregated result of customer preferences and the social and market environment, we exhibit how valued-ERGM models can be used to model directed and undirected product competition networks with non-binary link strengths. The method enables designers to estimate the major factors that affect customers' consideration and choice behavior, and which can help in predicting the strength of future market competition when a manufacturer changes some product attributes.

This work has three main contributions. First, we extend the newly developed valued-ERGM, which has traditionally been confined to social network modeling, to study competition between products. This network modeling approach enriches the knowledge base of product design modeling techniques. Second, by developing a procedure of weighted network construction, interpretation, and validation, we demonstrate that valued-ERGM models provide a better model than binary-ERGM, as measured by model fit and prediction accuracy for car competition. Third, this study is the first to study aggregated purchase preferences using a "directed" uni-dimensional network. The directed network we create is unique, as it encodes information from two stages of decision-making, both the final purchase decision as well as the items considered by the customers.

The case studies in this chapter show how network models are used to systematically analyze large real-world networks. For the first case study, which examines the co-consideration competition between 296 cars, we show that homophily effects, affecting the differences between two cars, are more important than the main effects in predicting link strength. Cars are generally found to compete for more with other cars from the same market segment, same brand origin, and similar price range. In the second case study, which focuses on the crossover SUV market, we analyze a network of 217 cars and find that cars that more people consider are also purchased more often.

# CHAPTER 4
# A FRAMEWORK OF INFORMATION RETRIEVAL AND SURVEY DESIGN FOR TWO-STAGE CUSTOMER PREFERENCE MODELING

## 4.1 Introduction

Designing a customer-favored product is critical to a company's success in a competitive market. Companies are particularly interested in what factors influence customer (one who purchases or receives a product or intends to do so) purchase behaviors and their relative importance. In the past decades, customer preference modeling has been a primary research method to answer these questions in both marketing science (Pescher & Spann, 2014; Stankevich, 2017) and the engineering design community. For example, customer preference modeling can provide designers with insights into identifying customer-preferred product features and how customers make tradeoffs among multiple attributes (Pescher & Spann, 2014; Sha, Wang, et al., 2017). Furthermore, research shows that a customer's decision-making process typically involves two stages during which the customer first forms a consideration set and second makes the final choice using different criteria (Shocker et al., 1991). The interest in customer preference modeling has primarily focused on two aspects: 1) to understand how product attributes influence customers' decision-making. For example, attempts have been made to model the impact of product design attributes on customer considerations and choices using customer-product network modeling (Bi et al., 2021; M. Wang, Chen, Huang, et al., 2016). 2) To understand the role of social influence in customers' decision-making (Argo, 2020), for example, using the data on customer-preferred product attributes before and after peer effects (Narayan et al., 2011) and demographic data from customers' social neighbors (Aral & Walker, 2011; Campbell & Lee, 1991). However, one major gap in current literature is that the impact of social influence and product attributes on customer purchase decisions are investigated separately. This is attributed to the limitations of data in two aspects. First, customers' social network data and the attribute data of their considered and purchased products are not collected simultaneously. Therefore, synthetic social network data has to be created when studying the social influence on customers' choices (L. He et al., 2014). Second, many datasets came from the private sector. Since those data often embed customer preferences, it is of high commercial value to enterprises, thereby cannot be shared publicly. Consequently, such limitations have affected the reproducibility and repeatability of many existing models.

To overcome these limitations, researchers must settle for the second-best to explore obtainable data sources, such as online product reviews, social media, and public customer survey data. Regarding the online review data, the reviews are typically generated by customers who have purchased the products (Lee & Bradlow, 2011), accessible via online stores' websites. Social media data are referred to the online content that customers or experts post on social network platforms such as Twitter or YouTube (Tuarob & Tucker, 2015). However, both types of data have minimal customer demographics, so customer reviews can not be associated with, yet, essential to customer preference modeling. Public customer survey data often includes a few products selected from a large pool of available products and can only support modeling studies with constrained information (Bao et al., 2020; Barnard et al., 2016). This study aims to develop a systematic approach that combines information retrieval and survey design in support of data collection for customer preference modeling that can address the limitations above. Specifically, we have made the following contributions: 1) we created a tool that can extract critical product features from customer reviews, integrating web scraping, text mining, and rule-based semi-supervised learning. 2) We developed a web-based survey platform that supports interactive information retrieving and virtual online shopping. 3) In the survey design, data quality assurance mechanisms, such as customer memory tests and attention check questions, were created and added. 4) The survey supports collecting customers' social network data and their preferences in a unified framework. 5) We designed the survey to support the data collection of both customers' considerations and choices. Thus, the data collected can be used in multi-stage choice modeling to study customers' consideration-then-choice behaviors. Our approach is demonstrated in the customer preference modeling of vacuum cleaners. To benefit a broader community, both the product and customer survey datasets will be made publicly available for researchers interested in customer preference modeling.

## 4.2 General framework of the information retrieval and survey design

In this section, we would propose a general framework in order to conduct such research and collect effective data. This framework can be used to collect customer data in different product markets, and it is capable to be used for a variety of customer-preference modeling. The unique information we are collecting includes customers' two-stage decision-making process as well as their social relations.

Figure 4.1 depicts an overview of the proposed information retrieval and survey design approach for two-stage customer preference modeling. It consists of four major modules and two outputs. Then we provide the description of each module.

**Module 1: Product database establishment** The main goal of Module 1 is to create a product database with basic product information, such as product model names and product attributes. This database acts as an input that is linked directly to the subsequent survey design modules. We first design a well-formatted SQL database. Then, a web scraping tool is used to collect product information, e.g., product image and attributes, etc., and customer review data from major electronic

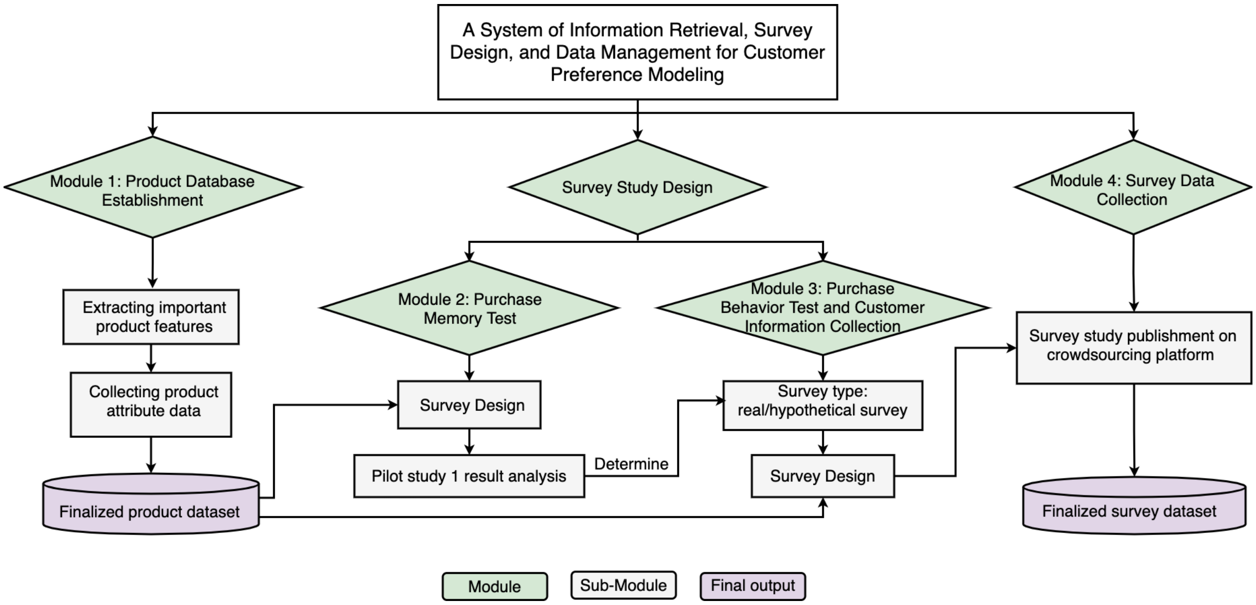

A System of Information Retrieval, Survey Design, and Data Management for Customer Preference Modeling
Module 1: Product Database Establishment
Survey Study Design
Module 4: Survey Data Collection
Extracting important product features
Module 2: Purchase Memory Test
Module 3: Purchase Behavior Test and Customer Information Collection
Survey study publishment on crowdsourcing platform
Collecting product attribute data
Survey Design
Survey type: real/hypothetical survey
Pilot study 1 result analysis
Determine
Survey Design
Finalized product dataset
Finalized survey dataset
Module
Sub-Module
Final output

Figure 4.1: The general framework of the information retrieval and survey design

retailers and department stores, e.g., Amazon, BestBuy, and Walmart. Next, utilizing text mining technology (e.g., a two-fold rules-based model (TF-RBM (Rana & Cheah, 2017)), we extract all the product attributes from scrapped customer reviews and allocate quantitative importance scores to each identified attribute based on its frequency of occurrence within the scraped reviews (Rai, 2012). The final list of critical attributes is determined by the rank of their importance scores and expert input. Finally, all of the collected data is organized and saved in the SQL database.

**Module 2: Purchase memory test** When taking a survey, the amount of detailed product information (e.g., the model name) a participant could memorize depends on how long the product was purchased. This leads to the idea of creating Module 2 to account for the memory bias across different participants. Therefore, to ensure the data quality, a purchase memory survey test, e.g., whether the customers who purchased a vacuum cleaner in the past one month, three months, or six months can remember their choices, is designed prior to the formal survey study. Once the memory test result is obtained, we use the test result to determine the type of survey, revealed or stated. In the revealed study, only the participants who actually purchased the product will be eligible to take the survey, and the data will be used for model revealed preference. Whereas in the stated study, the participants are required to complete the survey based on a virtual online shopping experience.

**Module 3: Purchase behaviour test and customer information collection** Module 3 focuses on the questionnaire design of the customer preference survey. We divide our questionnaire into three major parts to ensure that the collected data can support both the social influence and the consideration-then-choice behavior analyses. Part One is to collect participants' historical consideration and choice data, including the type of product they considered, the exact model they eventually purchased, and the top-rated attributes (features) that influenced their choice-making. In Part Two, we design questions to collect participants' social network data. This includes both their general social networks (GSN) as well as product-specific social networks (PSN) (Campbell & Lee, 1991). The GSN is a natural social relation network that captures the people with whom respondents communicate about important issues in their daily lives, such as their spouse, parents, and close friends. The PSN refers to the people with whom respondents have discussed product purchases, such as their coworkers who have endorsed their purchase, and they may or may not be from respondents' GSN. A person's PSN has the potential to influence their choice behaviors. Part Three focuses on gathering participants' personal information and user preferences. This includes their demographics, general preferences for household appliances, and product usage context. We use a variety of strategies to guarantee data quality (Bernard, 2013). These strategies are: 1) developing a product searching system to reduce participants' manual workload, thus improving the information retrieval accuracy; 2) setting attention check questions; 3) conducting both internal and external pilot studies; 4) implementing phase-in data collection and adjustment; and 5) incorporating experts' inputs and feedback from multiple disciplines including engineering design, social science, and psychological science.

**Module 4: Survey data collection** Module 4 is associated with two tasks: 1) designing a well-formatted and structured database that is advantageous for later data utilization, and 2) launching the survey on a crowdsourcing platform. The reputation of the crowdsourcing platform is essential because it directly influences the quality of the participants we can recruit. A platform with quality assurance mechanisms such as an AI-drive fraud detection system is always beneficial for us to collect high-quality data. Some popular platforms include MTurk, Prolific, and Cint. Once the data is collected, it is automatically saved in the SQL database.

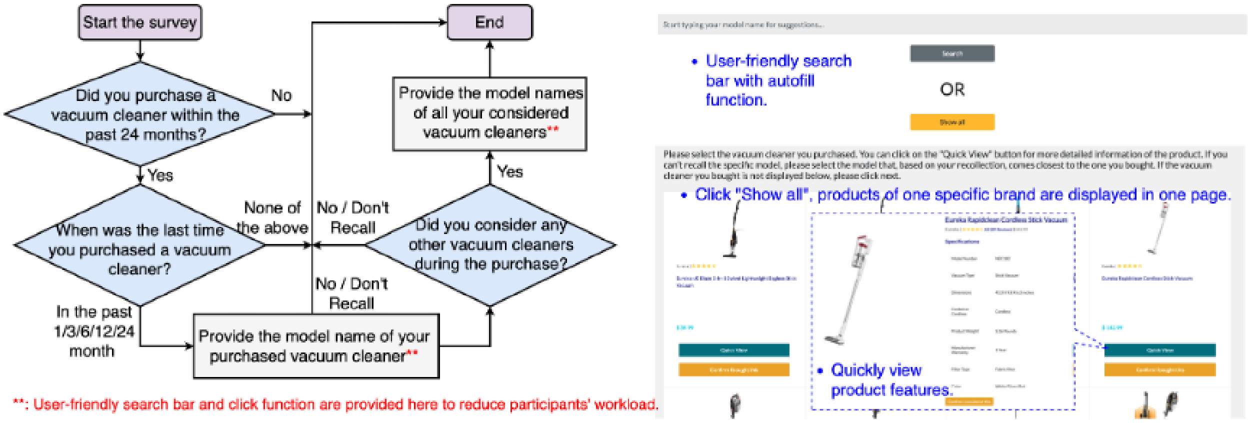
## 4.3 Survey design on household vacuum cleaner

In this study, we focus on a specific product market: household vacuum cleaners. There are several reasons for the selection: 1) it is a common household appliance with heterogeneous categories (e.g., upright, canister, robotic, etc.) and multiple competitors (e.g., Dyson, Shark, etc.) in the market; 2) it has a large market size with customers who have heterogeneous preferences on vacuum cleaners based on their demographics and usage context, and 3) its design attributes (features) play an important role in influencing customers' choices (Harmer et al., 2019), so the study on customer preference modeling shed light on design for market systems.

### 4.3.1 Vacuum cleaner data collection and attribute extraction

We scraped vacuum cleaner information and built the product database using the web crawling technique (Beautiful Soup and Selenium in Python). The household vacuum cleaners had been scrapped from the mainstream online shopping platforms in the US market (Amazon, Wayfair, Best Buy, Home Depot, and Walmart). Meanwhile, by scrapping the structured website, we collected the product information (product title, customer rating, SKU (stock-keeping unit)), features (list price, product dimension, weight, manufacture, brand, color, capacity, etc.), product description, and customer reviews. This study focused on five primary categories of vacuum cleaners - upright, canister, stick, handheld, and robotic vacuum cleaners. Data cleaning was performed to merge data from different sources, remove duplicated models and noises, and perform text mining to identify missing feature values. In the end, 1170 products with 26 features were collected in our final dataset.

In addition, we extracted product features from online customer reviews to determine the most important (most frequently mentioned) features to be included in the survey questions. We scrapped 60,000 reviews from Amazon (200 reviews for each product) and used a rule-based semi-supervised learning model for extracting features and sentiment/opinion associated with those features. For example, some feature-opinion pairs extracted from the reviews include “strong suction,” “heavyweight”, “annoying cord,” and “loud noise.” After obtaining candidate features from the opinion mining, unrelated features were pruned, and the rest features were ranked based on their frequencies in customer reviews. In the end, we identified 22 important product features based on the opinion mining results, including attributes such as price, product type, floor surface recommendation, suitable for pet hair, suction power, noise, power source, bag or bagless, cord or cordless, battery charge time, HEPA filter, warranty, brand, color, weight, dimensions, power, capacity, navigation system, voice control, remote controls (robotic vacuum cleaner specific attributes) and overall customer ratings.


Start the survey
Did you purchase a vacuum cleaner within the past 24 months?
No
Yes
When was the last time you purchased a vacuum cleaner?
In the past 1/3/6/12/24 month
None of the above
No / Don't Recall
Provide the model name of your purchased vacuum cleaner**
No / Don't Recall
Did you consider any other vacuum cleaners during the purchase?
Yes
Provide the model names of all your considered vacuum cleaners**
End
Start typing your model name for suggestions...
• User-friendly search bar with autofill function.
Search
OR
Show all
• Click "Show all", products of one specific brand are displayed in one page.
Eureka RapidClean Cordless Stick Vacuum
Specifications
Quick View
• Quickly view product features.

Please select the vacuum cleaner you purchased. You can click on the "Quick View" button for more detailed information of the product. If you can't recall the specific model, please select the model that, based on your recollection, comes closest to the one you bought. If the vacuum cleaner you bought is not displayed below, please click next.

**: User-friendly search bar and click function are provided here to reduce participants' workload.

(a) Survey questionnaire design for customer purchase memory test

(b) Survey web platform design for customer purchase memory test

Figure 4.2: Survey questionnaire and web platform design for customer purchase memory test

### 4.3.2 Customer purchase memory test

A pilot study was conducted to assess customers' abilities to remember their vacuum cleaner purchase decision-making over the past one month, three months, six months, twelve months, and 24 months to determine the appropriate threshold in soliciting participants. We first built a survey web for the test. The survey design logic and web interface examples are shown in Figure 4.2. To reduce participants' workload, a simulated online shopping system with features such as a user-interactive search bar and product preview was developed. As shown in Figure 4.2, we collected 30 samples for each period separately. Then, using those 30 samples, we calculated the proportion of participants who can recall the specific models they considered and purchased. Normally, if the ratio is greater than 50%, we consider customers' memory within that time period to be reliable.

The survey was conducted on the Cint platform from December 18 to December 21, 2020. Table 4.1 summarizes the actual collected sample size for the test. Because there were far fewer samples, the 24-month scenario was neglected in the proportion calculation. According to Figure 4.3, approximately 62% of customers who purchased a vacuum cleaner in the past three months can

Table 4.1: The sample size of the purchase memory test

| | In the past 1 month | In the past 3 months | In the past months | In the past 12 months | In the past 24 months |
|---|---|---|---|---|---|
| Number of people who have purchased a vacuum cleaner | 32 | 34 | 32 | 35 | 8 |

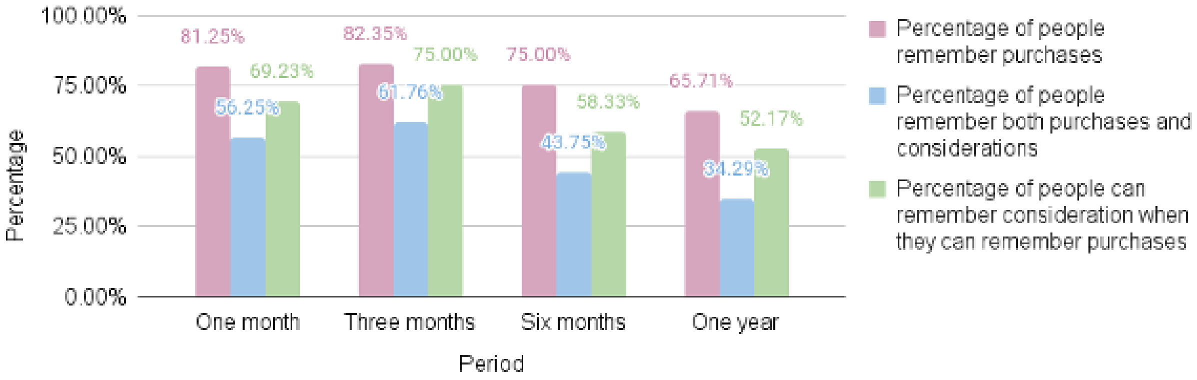


Figure 4.3: The ratio of participants who can recall the purchased or considered vacuum cleaners

remember their purchases and considerations, satisfying the 50% threshold. However, if we only focus on the customers who purchased vacuum cleaners within the past three months, we may not be able to collect enough samples for our following-up formal survey. Thus, we made a tradeoff by extending the period to the past six months because it has a high ratio of recall for purchase (75%); meanwhile, the ratio of recall for both purchase and consideration (43.75%) is still acceptable. So, in the formal study, only the customers who purchased the vacuum cleaner in the past six months were eligible to participate in the survey.

### 4.3.3 Vacuum cleaner customer survey questionnaire design

The customer purchase behavior test, as introduced in Section 4.2, consisted of three major parts. Part One employed the same simulated online shopping system to alleviate respondents' workload.

Furthermore, participants can rank the product features that influence their decision-making by dragging them from a list to the corresponding text boxes. The list contained all of the attributes identified by the feature selection algorithm introduced in Section4.2. In Part Two, participants were asked to provide at least one and up to five individuals' information in their general social networks (GSN) as well as all the ones with whom they had discussed the vacuum cleaner purchase. These individuals' demographic information and their contact frequencies with the respondents were also recorded. Part Three collected respondents' personal information and attributes, such as their own stated product preferences. To ensure the data quality, aside from the strategies mentioned in Section 4.2, other strategies employed include 1) setting filtering questions, e.g., did you purchase a vacuum cleaner in the past six months, so that only satisfied respondents can participate in this survey; 2) organizing questions by placing important questions first and less important questions last; 3) making questions mandatory to avoid missing data, i.e., participants could proceed the next stage of survey only after answering all the required questions. Lastly, similar to the purchase memory test, an associated survey website of the purchase behavior test was designed.

### 4.3.4 Survey data collection

We employed the Cint platform to launch our survey due to its established reputation. Additionally, we developed an SQL database on pgAdmin with a fine-tuned column sequence to ensure that all the respondents' answers could be structurally saved. Note that this database had been configured to communicate effectively with the survey website. To acquire more results, the survey was distributed to different groups, such as those who had recently purchased a vacuum cleaner or those who are interested in home decoration and home appliances. The survey was conducted over a two-month period, from April 25 to June 25, 2021, and a total of 1011 responses were received, with a completion rate of 15.35%. Meanwhile, to mirror the real market, a quota sampling technique

Table 4.2: Summary of key usage contexts of survey respondents

| Cleaning frequency | Frequency | Percentage (%) | Number of pets at home | Frequency | Percentage (%) |
|---|---|---|---|---|---|
| Every day | 343 | 33.93 | 0 | 197 | 19.49 |
| Every week | 630 | 62.31 | 1 - 3 | 731 | 72.31 |
| Every month | 34 | 3.36 | Over 3 | 93 | 8.21 |

(Sudman, 1966) was used to match the age distribution of the US census.

## 4.4 Descriptive analysis of the survey data

In this section, after cleaning and processing the raw data, we assessed the utility and quality of our data by performing a descriptive analysis of customers' two-stage decision-making processes and social network influence. We also constructed the unidimensional co-consideration and choice networks using our survey data for visualization, which shows the potential of our survey data in supporting customer preference modeling and engineering design.

### 4.4.1 Descriptive analysis of customer-product data in two-stage decision-making

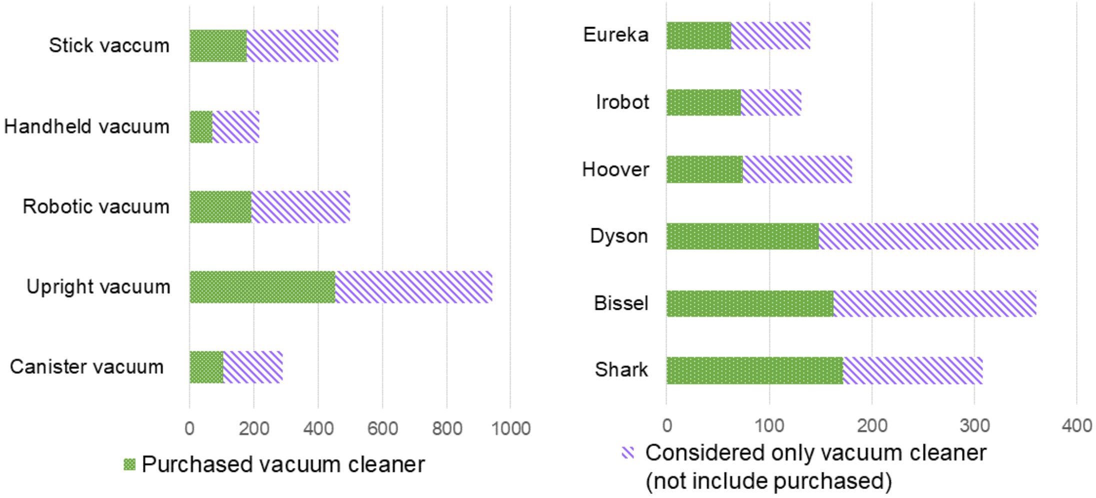
**Survey respondent demographics and usage context** From the demographic data, the average profile of respondents is male (56.87%), Caucasian (74.88%), 35-54 (29.48%), married (63.11%), retired (11.51%), with a bachelor's degree (36.80%), and with an annual household income of $40k - $70k (24.53%). The majority of respondents live in their own homes (76.55%), live in a single house (80.12%), have 6-10 rooms (55.59%), have stairs (65.18%). have multiple types of floors (70.43%), clean their home every week (62.31%), and have at least one pet (80.51%). Table 4.2 is a list of major usage contexts of survey respondents. (The detailed summarization and description of the data will be found in the appendix.)

**Considered and purchased vacuum cleaners** We collected information on the vacuum cleaners that respondents had considered and purchased as part of the study. Respondents reported 1011 vacuum cleaners purchased and another 1473 vacuum cleaners considered but not purchased. About 73.49% of respondents said they considered other vacuum cleaners before making their purchases, while 21.36% said they considered another vacuum cleaner, 28.19% said they considered another two vacuum cleaners, 19.99% said they considered more than three (up to six) vacuum cleaners.

The majority of vacuum cleaners that respondents have purchased (the solid green bar) and considered (the dashed purple bar) are shown in Figure 4.4(a). The total length of each bar indicates the popularity of each type of vacuum cleaner in customers' consideration, while the green bar reveals the popularity in customers' final choices. It's worth noting that upright vacuum cleaners are the most popular at both stages of consideration-then and choice. Figure 4.4(b) records the most popular brands that have been considered and purchased by respondents. It is noted that Dyson and Bissel are the most popular among respondents in the consideration, but Shark gains more popularity in the choice stage.

**The rank of importance for product attributes in two-stage decision-making** We have collected respondents' stated preferences regarding the most important features of vacuum cleaners in their decision-making process. Respondents were asked to pick and rank 3 - 5 of a vacuum cleaner's most important technical features in their consideration stage and choice stage. The importance of the attributes can be obtained by calculating the weighted sum of customer rankings, as shown in the following equations:

$$A = \sum (w_i \times c_i), for\ i = 1, 2, 3, 4, 5 \tag{4.1}$$



(a) Types distribution

(b) Brands distribution

Figure 4.4: Respondents' considered and purchased vacuum cleaners (a) types distribution and (b) top 6 brands distribution

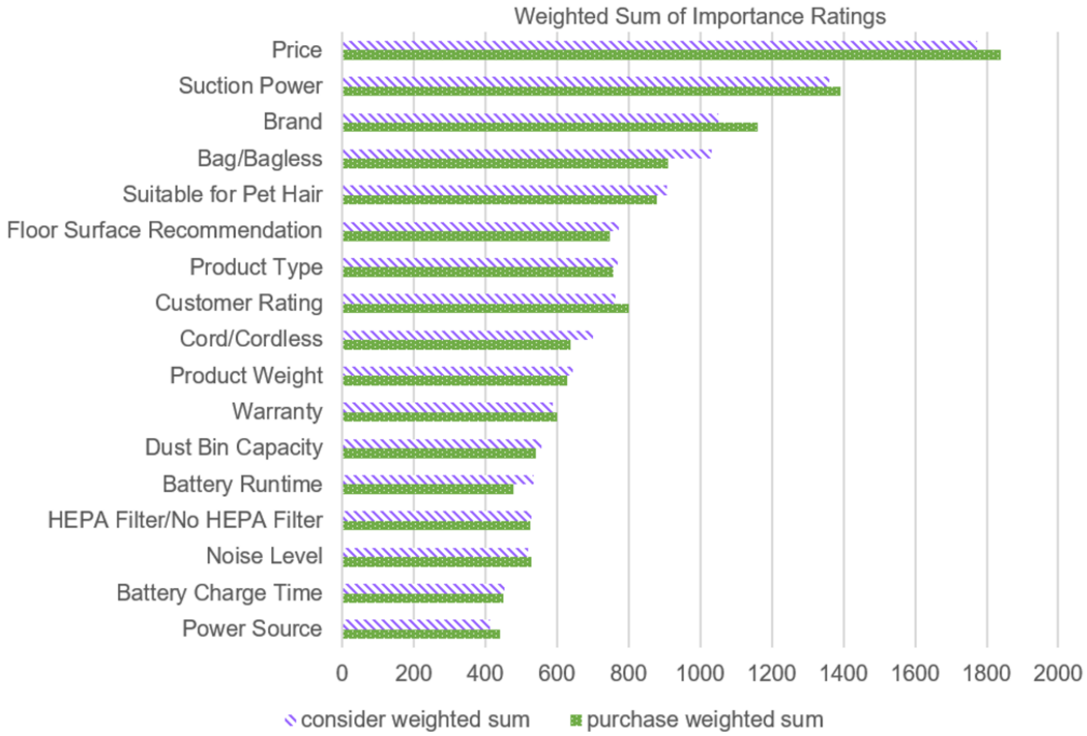


Figure 4.5: The rank of technical attributes based on the weighted sum of customers' importance rankings in their consideration stage and purchase stage

where $A$ is the weighted sum, $w$ is the ranking weight, $c$ is the count of the rank, and $i$ is the rank. We assign the ranking weight as 5,4,3,2,1 when the feature is rated as 1st, 2nd, 3rd, 4th, and 5th important. Figure 4.5 shows the important ranking of vacuum cleaner attributes based on the weighted sum of importance.

We can see the overall trends are consistent in consideration and choice stages, indicating price, suction power, and brand are the more important features in their decision-making process. Besides, there are some discrepancies between the two stages. For example, in the consideration stage, features such as product types, cord/cordless, bag or bagless, and floor surface recommendation are more important, while in the second stage, detailed and technical features such as price, suction power, and customer ratings are more important.

### 4.4.2 Social network influence analysis

In our survey, we asked respondents to name the people (up to 5 people) with whom they most frequently discuss important matters, as well as whether they had discussed their vacuum cleaner purchases with those people or anybody else (up to 5 people). In such a way, we investigate the respondents' general social network and vacuum cleaner-specific social network.

**General Social Network (GSN)** The respondents' general social network consists of people with whom they discuss important things in their daily lives. According to the results, respondents named 2.15 people on average, and the frequencies of naming a certain number of people are presented in Table 3 (the number of people in GSN). Among the people in their GSN, the most frequent relationships are with spouses (24.72%), friends (23.94%), and parents (12.18%). We also looked at the vacuum cleaners owned by the people in the respondents' GSN. It turns out that individuals with the same make and model as the respondents account for 31.99% of the total. 13.14% have the same make but different models, and 7.53% have the same type but different makes and models. The data is a preliminary indicator that GSN has an impact on repsondents' vacuum cleaner purchase.

**Vacuum Cleaner-Specific Social Network (VCSN)** We further investigate the individuals with whom the respondents have discussed vacuum cleaner purchases. While the respondents talked about their vacuum cleaner purchases with an average of 1.77 people in their GSN, they also stated they had discussed their purchases with an extra 0.42 people on average (the frequencies of the number of people in VCSN are shown in Table 4.3, GSN&VCSN, and VCSN only). Among the additional persons outside of a respondent's GSN, 19.85% are their friends, 17.49% are their acquaintances, 9.22% are their spouses, 7.57% are their neighbors, and 2.73% are salespersons.

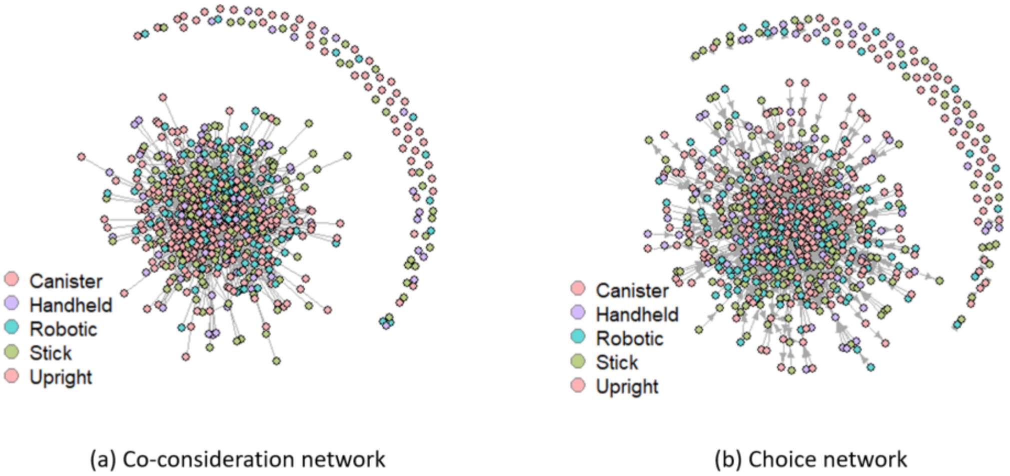
Table 4.3: The frequency of different numbers of the people in respondents' GSN and VCSN

| # of people in | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| GSN | 0.00% | 48.66% | 19.68% | 13.45% | 4.15% | 14.04% |
| GSN&VCSN | 8.11% | 48.76% | 20.08% | 11.67% | 3.76% | 7.62% |
| VCSN only | 73.69% | 19.49% | 3.07% | 1.09% | 0.40% | 2.27% |

According to the survey, people in respondents' vacuum cleaner-specific social network (VCSN) plays a vital role in their consideration and choice stages. For example, in their consideration stage, 42.55% of respondents say their VCSNs are very important (the highest among the five Likert scales). In their choice stage, 43.65% of respondents think their VCSNs are very important.

The data collected in this survey also includes the demographic information of the people in GSN and VCSN, the frequency of contact (which determines the strength of their relationship), and their personal viewpoints. All of the data we collect will be useful in understanding how social network influence affects customers' vacuum cleaner consideration and purchase decisions in future work.

**Co-consideration network and choice network construction** One important usage of the customer survey data is to build the customer-product networks based on the two-stage (consideration-then-choice) customers' decision-making process. As an illustration, we construct two simplified unidimensional networks, which only consist of the product nodes, to demonstrate the co-consideration and choice relationships among products. The undirected co-consideration network in Figure 4.6 (a) presents vacuum cleaner models as nodes, and the frequencies of two vacuum cleaners being co-considered by customers as links. The directed choice network in Figure 4.6 (b) presents the same set of nodes, while the directed links denote when two products are co-considered, which product between the two is more likely to be bought by customers.



(a) Co-consideration network

(b) Choice network

Figure 4.6: (a) Unidimensional co-consideration network and (b) choice network

In the co-consideration network, there are 672 unique vacuum cleaner models, while 63 products are isolated (which are not co-considered with others). The model “Dyson Upright Vacuum Cleaner, Ball Multi Floor 2, Yellow” is the most popular vacuum cleaner. It was co-considered with other vacuum cleaners by 46 times. The nodes’ average weighted degree is 6.45, which indicates that the vacuum cleaner models in our network are co-considered by 6.45 customers on average. In the choice network, there are 72 isolated items among the 672 vacuum cleaner models. The average weighted in-degree of the nodes is 1.79, implying that a vacuum cleaner is initially co-considered with other products before being picked by 1.79 consumers on average. The most popular purchased vacuum cleaner model is the same as the most considered vacuum cleaner model - “Dyson Upright Vacuum Cleaner, Ball Multi Floor 2, Yellow” (selected by 37 times). The product competition relationship can be represented by both the co-consideration network and the choice network, indicating customers’ aggregated preferences. Once these networks are constructed, more statistical analysis can follow to analyze and predict customer preferences.

## 4.5 Discussion

As discussed in Chapter 2, the collection of customer data presents a challenge in our study. Our objective is to investigate the influence of social factors on customers' two-stage decision-making behavior. In order to obtain pertinent data, we employ ego-specified information to capture the social influence experienced by each individual customer. However, it is important to acknowledge a limitation in our study. The design of the survey questions pertaining to customers' social influence could be enhanced. The current questions fail to adequately capture the products possessed by individuals who influence customers, thereby impeding the mathematical modeling of the extent to which social influence affects customers' decision-making. This limitation emphasizes the necessity for further refinement and fine-tuning of the survey questions to ensure a more comprehensive understanding of the dynamics involved in social influence on customer behavior.

## 4.6 Conclusion

In this study, we presented a systematic approach that combines information retrieval and survey design in support of data collection for customer preference modeling. This approach supports a systematic design of customer surveys that collect customers' social network and preference data in both the consideration and final choice stages. Therefore, the resulting datasets can support the study of a wide range of customer preference models, such as the social influence modeling and consideration-then-choice analysis, which can help product designers understand the feature's importance and make critical design decisions. Another merit of this study is the integration of state-of-the-art information retrieval techniques and survey design guidelines, including web scraping, text mining, SQL data management, and data quality assurance (e.g., purchase memory test). We have demonstrated how the approach works and how the techniques and guidelines are integrated using a case study on household vacuum cleaners. Our approach can be generally applied in collecting data on engineered products that are physical and discrete. We also conducted preliminary data analyses to assess the utility and quality of the obtained data. These data is available to the public for broader impact. In the next chapter, we will present an example that examines network-based customer preference modeling by utilizing the collected dataset.

# CHAPTER 5

# NETWORK-BASED ANALYSIS OF HETEROGENEOUS CONSIDERATION-THEN-CHOICE CUSTOMER PREFERENCES WITH MARKET SEGMENTATION

## 5.1 Introduction

A quantitative understanding of customer preferences plays a vital role in both product design and marketing strategy. In product design, it influences areas such as design attribute selection (Hoyle et al., 2009) and design optimization (Wassenaar & Chen, 2003). In marketing strategy, it guides initiatives such as the development of targeted advertising campaigns (John et al., 2018), pricing strategies (Draganska & Jain, 2006), and the identification of potential markets (Simons, 2014). Network-based models have been increasingly used to quantitatively analyze customer preferences and behavior (Sha, Wang, et al., 2017; M. Wang, Chen, Fu, et al., 2015). These models represent customers and products as nodes and their relations as edges in a network, allowing for the analysis of complex interactions and relationships between customers and products. One major advantage of network-based models over traditional utility-based choice modeling in customer preference modeling (K. Train, 1986) is their ability to handle both exogenous and endogenous attributes. Exogenous attributes can include product design features and customer attributes, while endogenous attributes may relate to effects from the market structure (Sha et al., 2019). In contrast, traditional utility-based choice modeling, which primarily focuses on customer and product attributes, often overlooks real-world factors like dependency on alternatives and the irrationality of customer decisions. Therefore, with their more comprehensive approach, network-based models provide a more flexible statistical inference framework than traditional methods.

Meanwhile, there is growing evidence in literature (Gaskin et al., 2007; Hauser et al., 2009; Shao, 2007; Shocker et al., 1991) on consumer research indicating that the complexity of customers' decision-making process consists of two different stages, consideration and choice, as shown in an illustrative example of vacuum cleaner purchase in Figure 5.1. In the consideration stage, customers make initial selections of products to form a consideration set. Then in the choice stage, customers evaluate the tradeoffs among the products in the consideration set to make a final choice. Our group has proposed a two-stage network-based modeling approach to study customers' consideration and choice behaviors (J. Fu et al., 2017). The result suggested that the factors influencing customers' consideration process differ from those shaping their final choice decisions. This study, however, analyzes the car market, where customers' considerations and choices are relatively constrained and centered around a single type of products. It is reasonable to assume that customers purchasing the same type of products share similar needs and preferences. For example, individuals who buy SUVs may appreciate their spaciousness and power. Once they have set their sights on an SUV, it is less likely for them to consider sedans, which are typically smaller in size and offer better fuel efficiency. The proposed network-based model in this study focuses on one type of product (i.e., sedans) and is able to effectively identify significant factors in customers' decision-making processes. In contrast, many other consumer product markets exhibit greater diversity, with customers choosing from a wide range of products. For instance, within the vacuum cleaner market, customers considering robotic vacuum cleaners might also evaluate the option of buying upright vacuum cleaners. In these markets, classifying customers solely by their purchased products is insufficient, as customers who make similar purchases still possess significantly distinct decision-making processes. Consequently, a systematic method for investigating heterogeneous customer preferences and segmenting customers is necessary in more diverse and mixed markets. This research plays a crucial role in understanding how product attributes influence customer behavior in highly varied product markets.

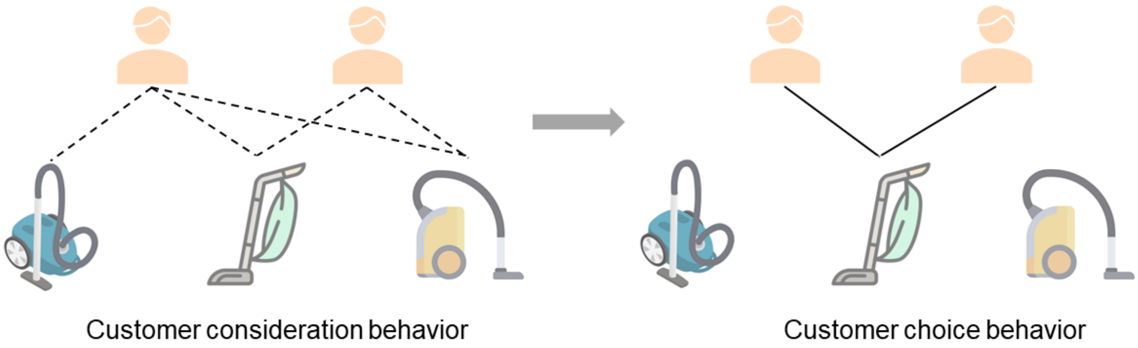


Figure 5.1: Two-stage consider-then-choose decision-making in an example of customers purchasing vacuum cleaners. The two-stage choice model assumes that each customer considers a subset of products first and then makes the final decisions.

In market research, segmentation is a commonly used method that simplifies the modeling of heterogeneous consumer preferences by categorizing customers into homogeneous groups or segments with similar characteristics and needs (Beane & Ennis, 1987; Goyat, 2011). It has been discovered that customer characteristics, including personal factors, psychological factors, and social factors, have a strong influence on their preferences (Kamakura et al., 1996). Correspondingly, typical market segmentation techniques include demographic segmentation (Lin, 2002)(such as age, gender, and income), psychographic segmentation (Lin, 2002) (such as people's lifestyles and personal viewpoints), behavioral segmentation (Susilo, 2016) (frequency of product usage), and need-based segmentation (Peltier & Schribrowsky, 1997)(usage context). Previous researches mainly apply these market segmentation techniques in conventional preference models, such as additive value functions (Liu et al., 2019) and discrete choice models (Kamakura et al., 1996). The integration of market segmentation into network-based preference modeling remains a research topic, and this study aims to address this gap in the literature. Particularly, we intend to answer the following three research questions:

- RQ1: How can we identify market segmentations based on customer characteristics?
- RQ2: Can network-based modeling that incorporates market segmentation lead to better results than using a single network model for the entire market?
- RQ3: What can we learn about the impact of product attributes on customers' two-stage consideration-then-choice decision-making process using market-segmentation-based network modeling?

To answer these research questions, we propose a market-segmentation-based network modeling approach to model heterogeneous customer preferences in a two-stage decision-making. In this approach, we first employ Joint Correspondent Analysis (JCA) to visualize the heterogeneity in customer preferences and how customer preferences are associated with different customer attributes (RQ1). Next, we cluster customers into different groups according to significant customers' attributes and construct consideration and choice bipartite networks for each group. Each bipartite network consists of customers and products, representing customers' two-stage decision-making process. Lastly, the Exponential Random Graph Model (ERGM), a statistical network modeling approach, is utilized to investigate the important factors that influence customer consideration and choices in different market segments. A single network model without market segmentation serves as the baseline to be compared with our proposed model (RQ2 and RQ3).

Our approach is demonstrated using the data from the vacuum cleaner customer survey, which was systematically designed to study multi-stage customer preference modeling in chapter 4. We chose the vacuum cleaner market as our research domain because it is a common household appliance with a diverse range of categories and a large customer base. The dataset contains 1,011 customer observations of 267 variables, including vacuum cleaner product attributes, customer purchase history (considered products and purchased products), and customer attributes (demographic attributes, usage context, and personal viewpoints).

## 5.2 Methodology

The methodology mainly contains two components: (1) customer segmentation informed by joint correspondence analysis, and (2) network construction and modeling. Figure 5.2 presents a comprehensive overview of the methodology.

### 5.2.1 Joint correspondence analysis and customer segmentations

First, we employ joint correspondence analysis (JCA) as an exploration tool to examine the relationship between customer characteristics and their preferences. These preferences are indicated by the products they consider. JCA facilitates in identifying the key customer attributes that drive their preferences, which are then utilized for clustering customers into distinct segment.

*Product community detection*

First, we identify customer preferences, represented by the product communities to which their considered products belong. In this context, product communities refer to groups of products that customers frequently co-consider. To identify product communities, We use their co-consideration relationships to construct a product association network. These relationships are derived from the survey data, where each customer reports the products they have considered. For example, in our vacuum cleaner market survey data, “Dyson Upright Vacuum Cleaner, Ball Multi Floor 2” and “Dirt Devil Razor Pet Bagless Upright Vacuum” have been co-considered by customers, resulting in a co-consideration link between them.

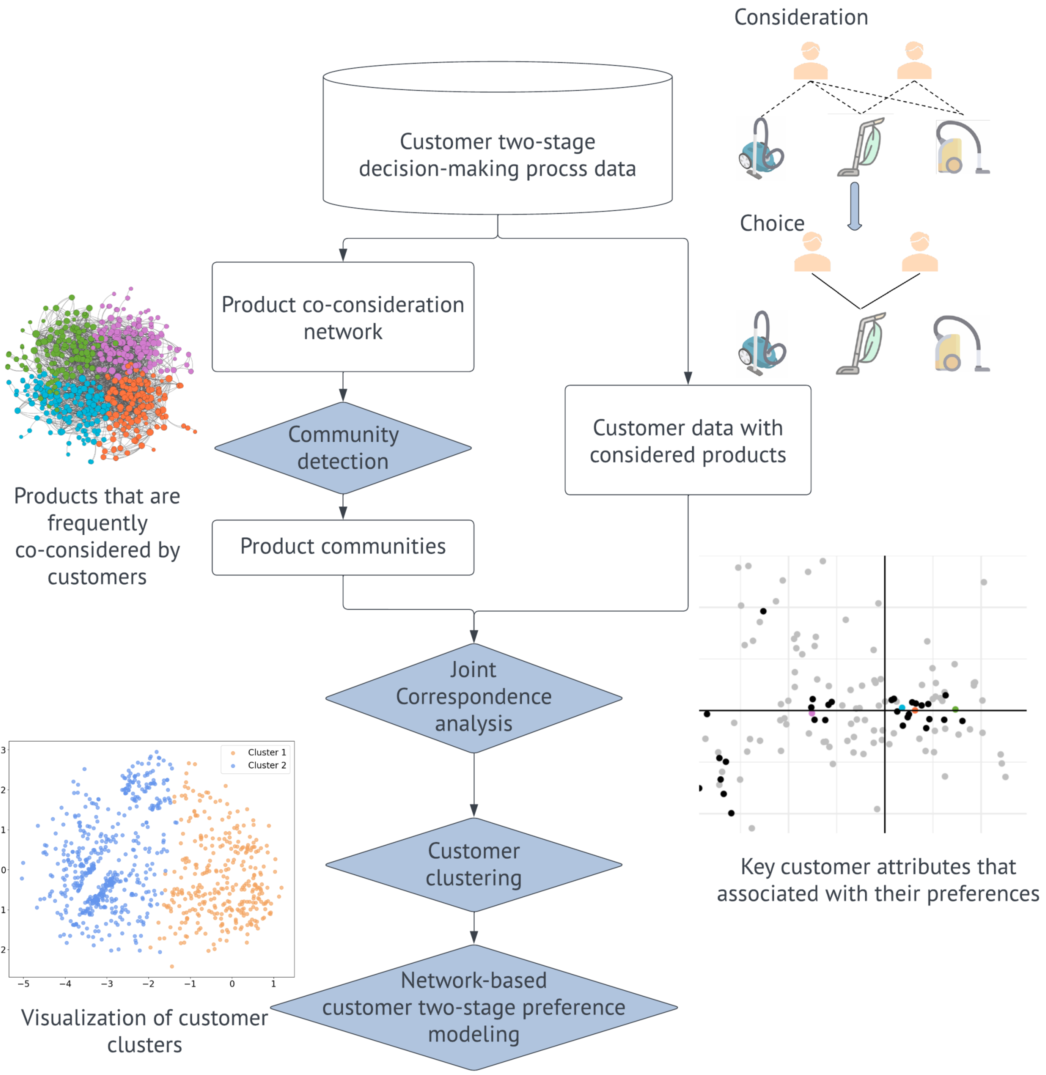


Figure 5.2: Framework of the methodology

In network theory, a community refers to a group of nodes within a network that are highly interconnected with one another, but relatively sparsely connected to nodes outside the community. Community detection algorithms partition a network into such groups. By applying these algorithms to a product co-consideration network, we can identify groups of products that are frequently co-considered. We then proceed to investigate the customer attributes associated with these product communities, which reflect their preferences.

The community detection algorithm we employ is based on a modularity-based optimization approach, as proposed by (M. E. J. Newman & Girvan, 2004). Modularity score quantifies the extent of connectivity within communities relative to what would be expected in a random network. Different detection results yield varying modularity scores, and by optimizing the modularity score, we can obtain a clearer separation of communities within the network. Many algorithms automatically detect the number of communities by optimizing the modularity score. However, in less dense networks, this approach may return a large number of communities, making it challenging to derive insights from the result of community detection. To address this, we can use the elbow rule to pre-define an appropriate number of communities based on a reasonably high modularity score (generally, a value above 0.3 indicates significant community structure in a network (Clauset et al., 2004)). The elbow rule (Ketchen & Shook, 1996) is a graphical method that plots the modularity score against the number of communities. The optimal number is usually at the point where the score starts to level off, resembling an elbow shape. To apply the elbow rule, we first perform community detection on the network for a range of community numbers (e.g., from 2 to 20). Then, for each community number, calculate the modularity score and plot it against the number of communities. The optimal number of communities is estimated by the point on the plot where the modularity score starts to level off, i.e., the ”elbow” of the curve.

In our study, we seek to find out the product communities, the groups of products that are frequently considered together by customers and thus are characterized by high levels of competition among the products within them. Instead of using product categories (such as upright vacuum cleaners, stick vacuum cleaners, and robotic vacuum cleaners) to represent product types, we choose product communities because they better capture customer preferences by more accurately reflecting the market structure.

### *Joint correspondence analysis*

Joint correspondence analysis (JCA) is a statistical technique used to analyze and visualize the relationships between multiple categorical variables. Originating from correspondence analysis (CA) – a method for analyzing two-way contingency tables based on the singular value decomposition of a matrix of a correspondence weight matrix – JCA extends CA to accommodate the analysis of multiple contingency tables or joint distributions (Greenacre & Blasius, 2006). This extension provides a more comprehensive view of the relationships by enabling the analysis of multiple variables simultaneously. Fundamentally, JCA is a dimension reduction method that enables visualization of a data matrix in a low-dimensional subspace.

Widely employed in marketing research, CA (or JCA) helps study the relationship between customers' preferences and their characteristics (Hoffman & Franke, 1986). For example, CA was used to identify customer groups with similar purchasing patterns (Beldona et al., 2005) and determine products that are most likely to be purchased together (Hoffman & Franke, 1986). Additionally, the technique has been employed in social sciences to investigate the connection between demographics, attitudes, and behaviors (de Nooy, 2003). In network-based customer preference modeling (M. Wang, Chen, Huang, et al., 2016), JCA was introduced as a multivariate approach that represents data graphically. This offers a visual understanding of the connections between product consideration sets and their relations with customer attributes.

After detecting product communities in the product co-consideration network, we employ JCA to explore the association between customer characteristics and their preferences for vacuum cleaners, which was represented by the product community that their preferred product belongs to. We consider a range of customer attributes, including demographic attributes, usage context attributes, and personal viewpoints about products. When using JCA, we consider $N$ customers' consideration observations, each associated with $x_1, \ldots, x_q$ categorical variables, which include both customer attributes and the product communities they have considered. Each $x_j$ has $L_j$ levels (the number of levels). From this data, we can generate a binary indicator matrix $\boldsymbol{Z}^{(j)}$ of dimension $N \times L_j$ for each categorical variable $x_j$, such that $\boldsymbol{Z}^{(j)}_{il} = 1$ if and only if $x_{ij} = l$, where $l$ is a level of $x_j$. These individual matrices $\boldsymbol{Z}^{(j)}$ can then be concatenated to form a large indicator matrix $\boldsymbol{Z}$ with dimensions $N \times J$, where $J = L_1 + ... + L_q$ represents the total number of categorical levels for all input variables $x$. Table 5.1 displays an example of the indicator matrix $\boldsymbol{Z}$ with three customers (with six customer consideration observations) in rows and four categorical variables in columns, including the product community variable $x_1$ that list the product community for customers to consider. Because the indicator matrix $\boldsymbol{Z}$ might consume significant memory resources when dealing with a large number of respondents and categorical levels, JCA operates on the Burt matrix $\boldsymbol{B} = \boldsymbol{Z}'\boldsymbol{Z}$, defined as the cross-tabulation of all categorical levels. Using the Burt matrix $\boldsymbol{B}$, the column coordinates relative to the principal axis can be calculated through Singular Value Decomposition (SVD), and JCA's corrected inertia is obtained by iteratively updating the solution.

Table 5.1: Indicator matrix in Joint Correspondence Analysis, with observations of customers' considerations as row entries, and considered product and customer attributes as column entries

| | Considered product community $x_1$ | | | Income $x_2$ | | | Pet $x_3$ | | Quality is important $x_4$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Com. 1 | Com. 2 | Com. 3 | Low | Mid | High | Yes | No | Agree | Neutral | Disagree |
| Customer 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| Customer 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| Customer 2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| Customer 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| Customer 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| Customer 3 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |

JCA transforms high-dimensional datasets into lower-dimensional spaces. By projecting variable levels into this space, a JCA plot reveals relationships between customer attributes and considered product communities. To identify the most influential customer features for subsequent customer segmentation, we mainly focus on two ways of interpretation:

- **Proximity between points**: In the plot, each customer attribute level and their considered product communities are represented by a point. Closely positioned points indicate a stronger association between the corresponding categories (either customer attributes or product communities). When two product communities are proximal, it suggests that they are often considered by customers with similar profiles. Two customer attributes are closer if they often appear together for specific customers. Last, when a product community and customer attributes are located near each other, it implies that customers who consider products in that community frequently possess those attributes. In our analysis, we focus on identifying customer attributes that are closer to each product community. This indicates a strong preference for certain products among customers with those attributes.
- **Column inertia**: Column inertia is a measure of the variation or dispersion of category levels in the multidimensional space. It quantifies the “distance” between each category level and the average or centroid for the respective category in the space defined by JCA. In simpler terms, it conveys how different or distinct a category level is compared to the average of all category levels in the same category. Higher inertia values signify that a category level is more distinct from the average, while lower values indicate its proximity to the average. In our case, we believe that a higher column inertia value suggests that a customer attribute level is more diverse or different from the average, indicating the potential significance in shaping customer preferences of the corresponding attributes.

Additionally, we take into account other findings obtained from the JCA analysis. Firstly, we examine the **explained variance in the lower dimensional space**. This metric indicates the percentage of the overall variation in the data that can be accounted for by the first two dimensions. A higher value signifies a better representation of the data variation by these dimensions. Furthermore, we investigate the concept of **opposite quadrants**. In the plot, product communities positioned in opposite quadrants tend to exhibit negative relationships. This implies that they are preferred by distinct customer segments or have divergent associations with customer attributes.

### *Customer segmentation*

Using key features associated with customer preferences identified through JCA analysis, we can categorize customers into distinct groups with unsupervised clustering. The customer clustering process includes selecting an appropriate clustering algorithm, determining the optimal number of clusters, and validating the clustering results with visualization and statistical tests, which evaluate differences between customer groups.

There are various clustering algorithms, including k-modes, hierarchical clustering, and density-based spatial clustering of applications with noise (DBSCAN). Different algorithms may be more suitable for different use cases. Moreover, different algorithms utilize distinct distance metrics, which are determined by the type of variables. For continuous variables, the k-means method calculates the Euclidean distance between observations. In contrast, for categorical and nominal variables that lack inherent order or rank, the k-modes method computes the Hamming distance, which measures the dissimilarity between categorical variables. For variables that are ordinal, possessing a natural order such as education level and personal viewpoints, it is crucial to maintain the order of different levels, which can be measured by Euclidean distance. However, the distance between groups may not be equal, making Euclidean distance potentially unsuitable. Consequently, selecting the most appropriate algorithm involves a trial-and-error process.

The optimal number of clusters can be determined using silhouette scores, a metric that measures the similarity within clusters and the dissimilarity between clusters. Higher silhouette scores indicate a greater degree of similarity among customer attributes within the same cluster.

Once the most suitable clustering technique and the optimal number of clusters are determined, we use dimensionality reduction methods, such as t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP), to visualize the clustered customer groups in lower-dimensional spaces and assess the clustering results. In addition to visualization, Chi-squared tests are conducted to further evaluate the significance of differences between the customer groups.

This comprehensive segmentation methodology is utilized to provide a deeper understanding of customer heterogeneity, enabling the development of tailored marketing and engagement strategies for each identified customer group. By addressing the unique needs and preferences of each group, we anticipate more accurate and effective results in subsequent stages of the study.

### 5.2.2 Two-stage customer decision-making process network modeling

In the second part of this study, bipartite customer-product networks are constructed, modeled, and interpreted for the different customer groups. These networks aim to reveal the underlying patterns of customer preferences among different groups and their interactions with diverse products. Simultaneously, the network-based model can uncover the underlying market structures. To further validate the model, a portion of the customers' considerations and choices is withheld during the modeling process. We calculate our model's accuracy on these withheld preferences and compare it to that of a benchmark model that doesn't take market segmentation into account.

### *Construction of consideration and choice networks*

When representing customer-product relationships with bipartite networks, customers and products are modeled as two distinct types of nodes. The considerations and choices of customers are depicted as different types of links. The customer decision-making process can be modeled as a probability of forming consideration or choice links between the two types of nodes. In the first stage, consideration links within the bipartite network signify customers' considerations among all available products. In the second stage, choice links represent the final purchase decision made by customers among the products being considered, conditional on the consideration set established in stage one.

In this study, we aim to integrate the customer segmentation results with the two-stage bipartite network models by constructing distinct network models for each customer segment. By doing so, we can leverage the insights from the customer segmentation analysis to better understand and predict the customer-product interactions within the bipartite network.

### *Statistical network modeling*

Recent advances in Exponential Random Graph Models (ERGMs) have offered a unified and flexible statistical inference framework for network analysis (Handcock et al., 2015), the formula of ERGM (Equation 2.5) was introduced in chapter 2. Bipartite ERGMs are a type of ERGM specifically designed for modeling two-level network structures (P. Wang et al., 2013). They follow the same model structure as defined in Equation 2.5, except that $g(y)$ must represent the network statistics specific to bipartite networks. Unlike one-mode networks, which consist of a single layer of nodes (e.g., social influence network and co-consideration network), bipartite networks are composed of links connecting two layers (two types of nodes). In this work, there are two types of links in two separate bipartite networks: consideration links or choice links.

The estimated parameters $\theta$ in the bipartite ERGM can be used to infer the effects of product features and market structure (of corresponding $g(y)$) on customer considerations and choices. By analyzing these effects, we can gain insights into the factors that drive customer decision-making and develop tailored marketing and engagement strategies to cater to the specific needs and preferences of each customer group.

### *Modeling setting for consideration-then-choice stages*

In modeling the two-stage decision-making process using the ERGM framework, we implement a constraint to ensure that only product nodes connected during the consideration stage can be linked to customers in the choice stage. By doing so, we accurately represent the process where customers choose products exclusively within their consideration sets, effectively capturing the nuances of their decision-making behavior.

### *Network simulation for prediction*

In addition to interpreting the important features in different customer segments, we also aim to validate the market-based-segmentation model performance by investigating the predictability of the model. The prediction capability serves as a validation metric for the effectiveness of the market segmentation compared to a benchmark model and could potentially be generalized to network-based choice modeling predictions.

To make predictions for the testing customers, we employ a method that treats all of the links between testing customers and available products as “missing links” (C. Wang et al., 2016). During the model estimation process, where we use observed network to calculate its parameters, these missing links remain absent and do not contribute to the calculation. Once the estimated parameters are obtained, we proceed to simulate the entire bipartite network, treating the connections between training customers and their considerations or choices as “observed link”, which remain unchanged during the simulation process. The simulation calculates the probability of all missing links based on the estimated ERGM parameters. The outcomes of this simulation serve as predictions for the customers’ considerations and choices.

By evaluating the prediction capability of our model, we can assess the effectiveness of customer segmentation and network-based choice modeling. Furthermore, this analysis allows us to validate and refine our approach, enhancing the accuracy of future predictions and providing valuable insights for targeted marketing and engagement strategies.

## 5.3 Case study: vacuum cleaner market

### 5.3.1 Data source

We demonstrate our methodology using customer survey data collected from the household vacuum cleaner market. This data, gathered through a questionnaire, includes information on 1,011 participants’ consideration and choice decisions, personal information, and product feature preferences. In addition, we obtained data on vacuum cleaner product features by web scraping on an online shopping website. In this study, we focus on the customer attributes (i.e., demographic attributes, usage context, and personal viewpoints about vacuum cleaners), their considered and purchased products, and product attributes.

### 5.3.2 Joint correspondence analysis and customer segmentations

*Product community detection*

The co-consideration network analysis reveals the presence of isolated nodes that are either not co-considered by other customers or only co-considered within 2-3 products. Since these nodes represented very small communities with low interest, they are excluded from the analysis. To ensure meaningful analysis, only the largest connected component in the co-consideration network is retained. In this context, the largest connected component refers to the biggest subset of nodes in the co-consideration network that are interconnected, based on which we conduct product community detection. We think that in the whole market with more extensive data collection, there won't be many isolated products. This step removes 6 percent of the products, and leaves 528 products for community detection.

In the product community detection process, the Spinglass algorithm (Reichardt & Bornholdt, 2006) is a flexible approach that allows users to predefine the number of communities. Additionally, it provides superior modularity scores compared to other algorithms with similar functionality. In this study, we apply the algorithm with elbow rules to determine the optimum community size, which is found to be 4 (Figure 5.3).

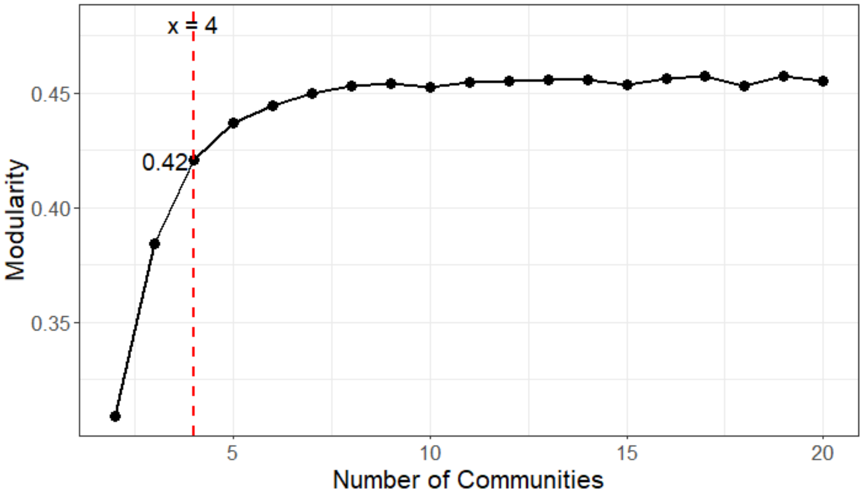


Figure 5.3: Using elbow rule to determine the number of communities

In Figure 5.4 and Table 5.2, we provide the community detection result and summarize the key characteristic with the mean value and standard deviation of each community. On average, Community 1 contains the most expensive vacuum cleaners, and the most dominant type is robotic vacuum cleaners, which may be summarized as “high-tech and expensive”. Community 2 contains the vacuum cleaners with lowest price and the most dominant type is upright vacuum cleaners, which can be named as “traditional and affordable”. In Community 3, the highest suction power is observed, which is represented by the attribute ”strong suction power.” Finally, Community 4 is dominated by lighter-weight stick vacuums, which are described as ”innovative and portable.” We represent customers’ preferences by the product communities of the products they have considered. This representation is effective because each product community is composed of closely interconnected products, which are often considered together by customers with similar tastes. Consequently, the specific product community associated with a customer’s considered products can provide a general indication of that customer’s preferences.

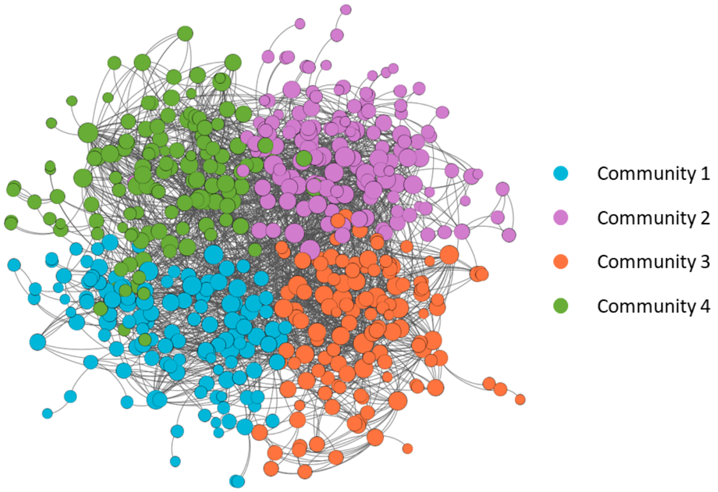


Figure 5.4: Community detection results of product co-consideration network

Table 5.2: Product community detection and characteristics (mean value and (standard deviation)) of each community

| Communities | Descriptions | Dominant Type | Suction power (rating) | Price (dollar) | Weight (lb.) | Representative models |
|---|---|---|---|---|---|---|
| Community 1 | High-tech and expensive | Robotic | 2.62 (1.25) | 296.12 (205.22) | 8.73 (5.09) | |
| Community 2 | Traditional and affordable | Upright | 2.90 (1.20) | 179.37 (166.19) | 11.68 (6.36) | |
| Community 3 | Strong suction power | Upright | 2.99 (1.34) | 259.85 (218.60) | 10.20 (6.78) | |
| Community 4 | Innovative and portable | Stick | 2.75 (1.22) | 228.47 (201.72) | 8.77 (5.76) | |

*Visualization of JCA results of customer attributes and product communities*

In the Joint Correspondence Analysis (JCA), customers' considered product communities and their associated characteristics are examined. We evaluated 31 customer characteristics related to demographic attributes, usage contexts, and personal opinions about vacuum cleaners, which comprised a total of 152 distinct category levels. These levels and product communities are depicted in Figure 5.5. Specifically, four separate product communities are denoted by diamond shapes in different colors. Important feature levels are represented as black dots, while less important ones are indicated by gray dots. Feature importance is determined by considering the **proximity** to the four product communities and the value of a **feature level's inertia**.

Feature levels with higher inertia values account for a greater variance within the data and deviate more significantly from the average, suggesting that the corresponding features may play a crucial role in differentiating customers. In Figure 5.5, these high-inertia features are labeled with their respective level names. Based on the JCA results, the top 10 features with the highest inertia values were selected, which include: PV5 (disagree with the importance of styling), PV4 (disagree with the environmentally friendly feature), Occupation (Retired), Own house (0), PV14 (disagree with the highest quality being important), Age (6), PV13 (disagree with advocating for a favorite brand), PV1 (neutral towards innovation), PV6 (disagree with the importance of energy efficiency), and Have house cleaner (Yes).

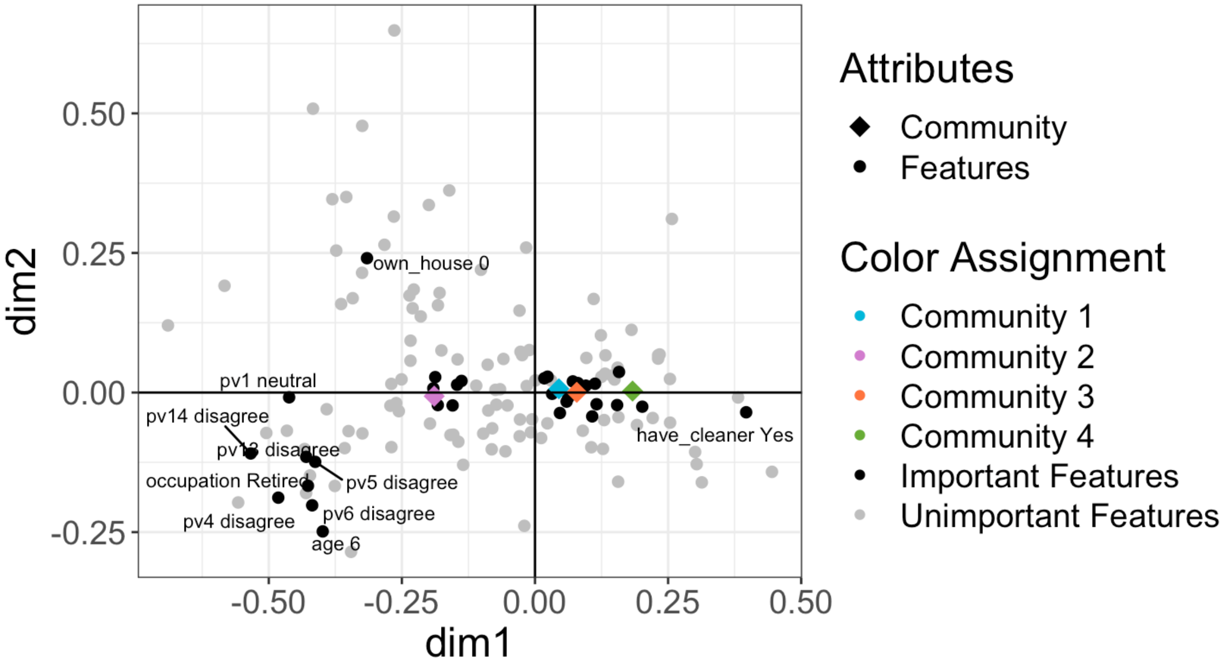
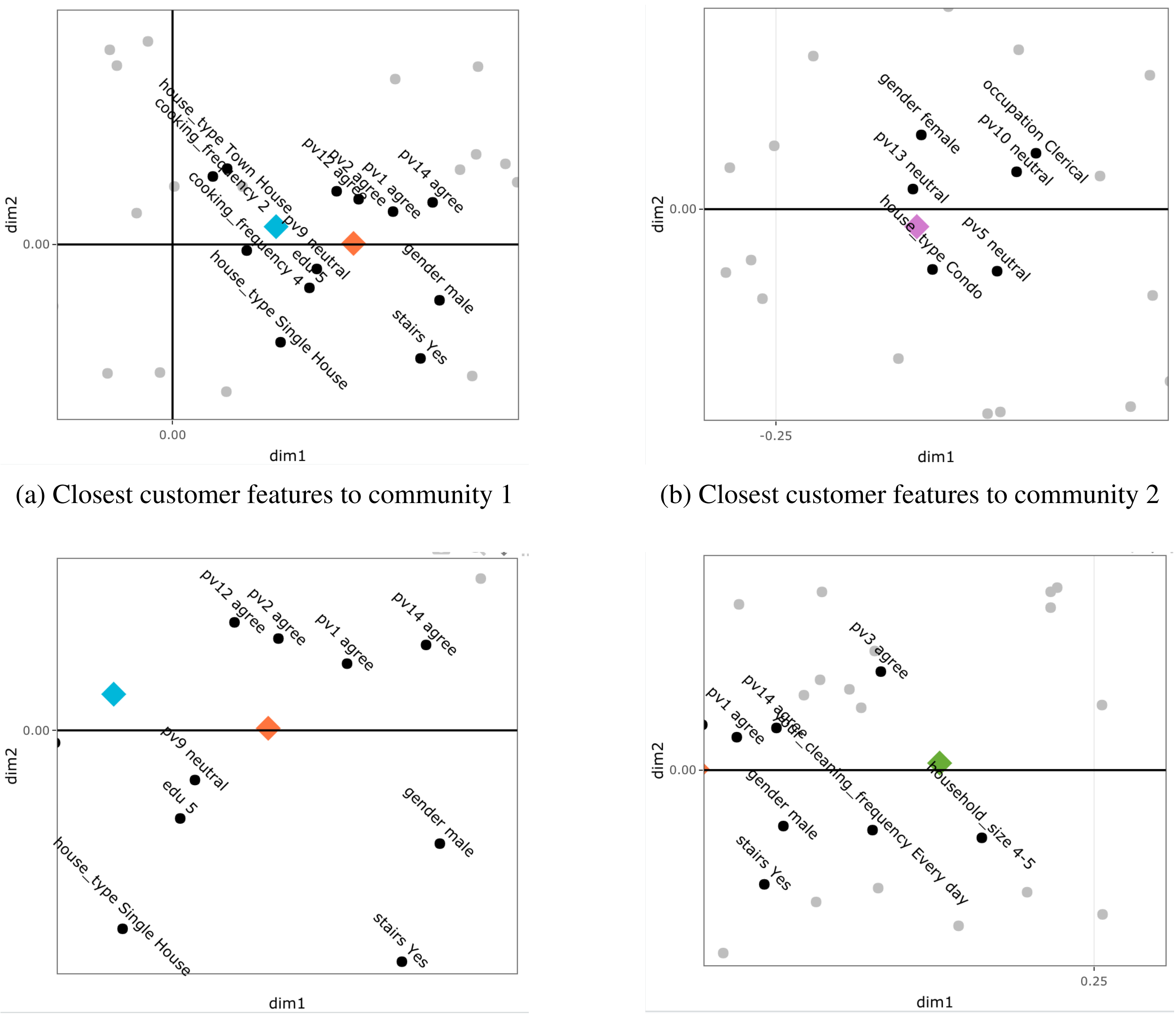


Figure 5.5: Joint correspondence analysis based on vacuum cleaner communities and customer attributes. The feature levels with high inertial are labeled with text. Product community 2 falls in the region of $dim1 < 0$, and product communities 1, 3, and 4, fall in the region of $dim1 > 0$

We further investigate the proximity between each product community and customer feature levels. The enlarged JCA plots for each community are displayed in Figure 5.6. The closeness between a product community and customer features indicates that customers with a particular feature often consider products from that community. For product community 1 (high-tech and expensive), it is strongly associated with customer features such as house type (townhouse, single house), cooking frequency (3-4 days a week, every day), bachelor's degree (Edu 5), tech enthusiasts (PV2 agree), and impressionable (PV12 agree). Product community 2 (traditional and affordable) is highly associated with customer features including occupation (clerical), gender (female), house type (condo), neutral opinions on styling (PV5), after-sale service (PV10), and advocacy (PV13). Product community 3 (strong suction power) is closely related to house type (single house), stairs (yes), gender (male), innovation enthusiasts (PV1 agree), tech enthusiasts (PV2 agree), impressionable customers (PV12 agree), and quality-conscious customers (PV14 agree). Lastly, product community 4 (innovative and portable) is strongly associated with features such as cleaning frequency (every day), house size (4-5), lifestyle-conscious customers (PV3 agree), and brand advocates (PV13 agree).

Based on both proximity and column inertia, we identify strong candidate features for customer segmentation, including demographic attributes (Education, Gender, Household size, Age), usage context (House type, cleaning frequency, cooking frequency, Own house, Have house cleaner), and personal viewpoints: innovation (PV1), tech enthusiasm (PV2), lifestyle compatibility (PV3), environmental friendliness (PV4), styling (PV5), energy efficiency (PV6), long product life (PV9), after-sale service (PV10), impressionability (PV12), brand advocacy (PV13), and quality-consciousness (PV14). These features provide valuable insights for effectively segmenting customers.

Moreover, it is evident that communities 2 and 4 are more distinguishable than communities 1 and 3. Also, Community 2 is located in the opposite quadrant to Communities 1, 3, and 4. This observation suggests that customers with a strong preference for community 2 products have distinct preferences compared to those who favor products in communities 1, 3, and 4. Additionally, the first two dimensions account for 60% of the data variance, indicating that these dimensions have



(a) Closest customer features to community 1

(b) Closest customer features to community 2

(c) Closest customer features to community 3

(d) Closest customer features to community 4

Figure 5.6: Customer features that have a strong association with each product community (proximity)

effectively captured a substantial amount of information present in the raw customer dataset.

In summary, the JCA plot shows that there are at least two major categories of customers, separated by $dim\ 1 > 0$ and $dim\ 1 < 0$, respectively. Their preferences significantly differ as they opt for vacuum cleaners associated with distinctive product communities. Utilizing the JCA results (proximity and column inertia), we have identified candidate features related to customer preferences that can be employed for effective customer segmentation.

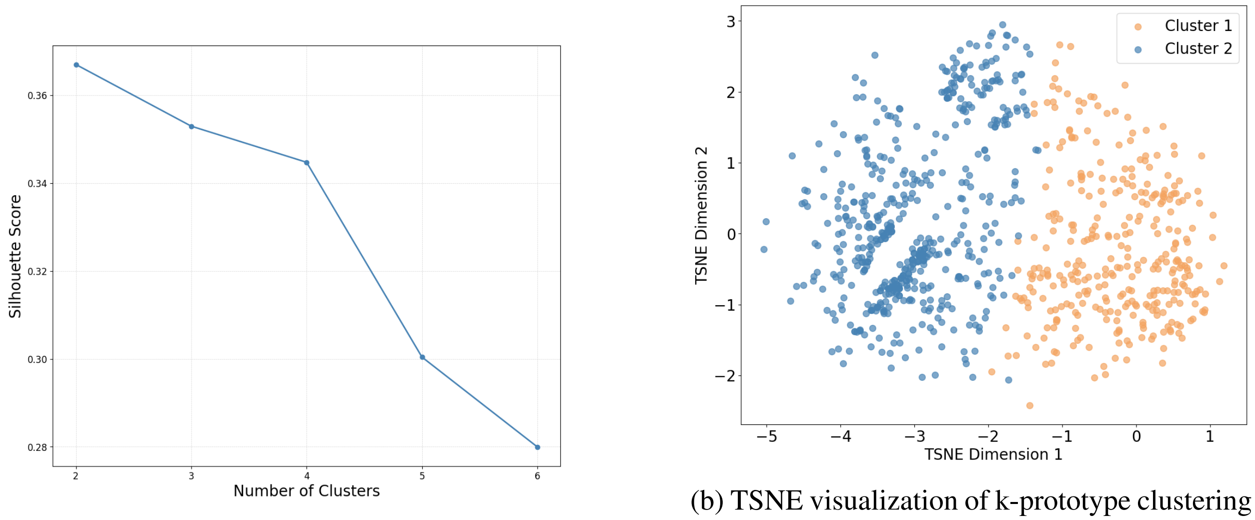
*Customer segmentation results*

Drawing upon the JCA results, the analysis utilizes a comprehensive set of 20 features spanning demographic attributes, usage context attributes, and personal viewpoints attributes for customer segmentation.

The selected customer attributes are all categorical, which can be further classified into nominal variables (e.g., gender and house type) and ordinal variables (e.g., education level, household size, and personal viewpoints). We use k-prototype clustering algorithm to cluster the customers based on these attributes. The algorithm effectively handles mixed data types by combining the k-means and k-modes algorithms. In this particular case, the Hamming distance is employed for nominal variables, while the Euclidean distance is used for ordinal variables. This approach, which takes into account the inherent order present in ordinal variables, yields better results in terms of silhouette score ignores this order.

To determine the optimal number of clusters, a silhouette score analysis is conducted. As depicted in Figure 5.7a, the silhouette score varies with the number of clusters, and the highest score is obtained with two clusters. This suggests that the clustering solution with two clusters maximizes the similarity within clusters while maximizing the dissimilarity across clusters. Using the k-prototype algorithm to classify the customers into two groups, we obtain 331 customers in cluster 1 and 540 customers in cluster 2. The clustering results are visualized with a t-SNE plot, which reduces the customers' attributes to a lower-dimensional space and reflects their similarity through the proximity of the points on the map, as shown in Figure 5.7b.

In addition to visual plots, we use chi-square tests to further confirm that the two customer

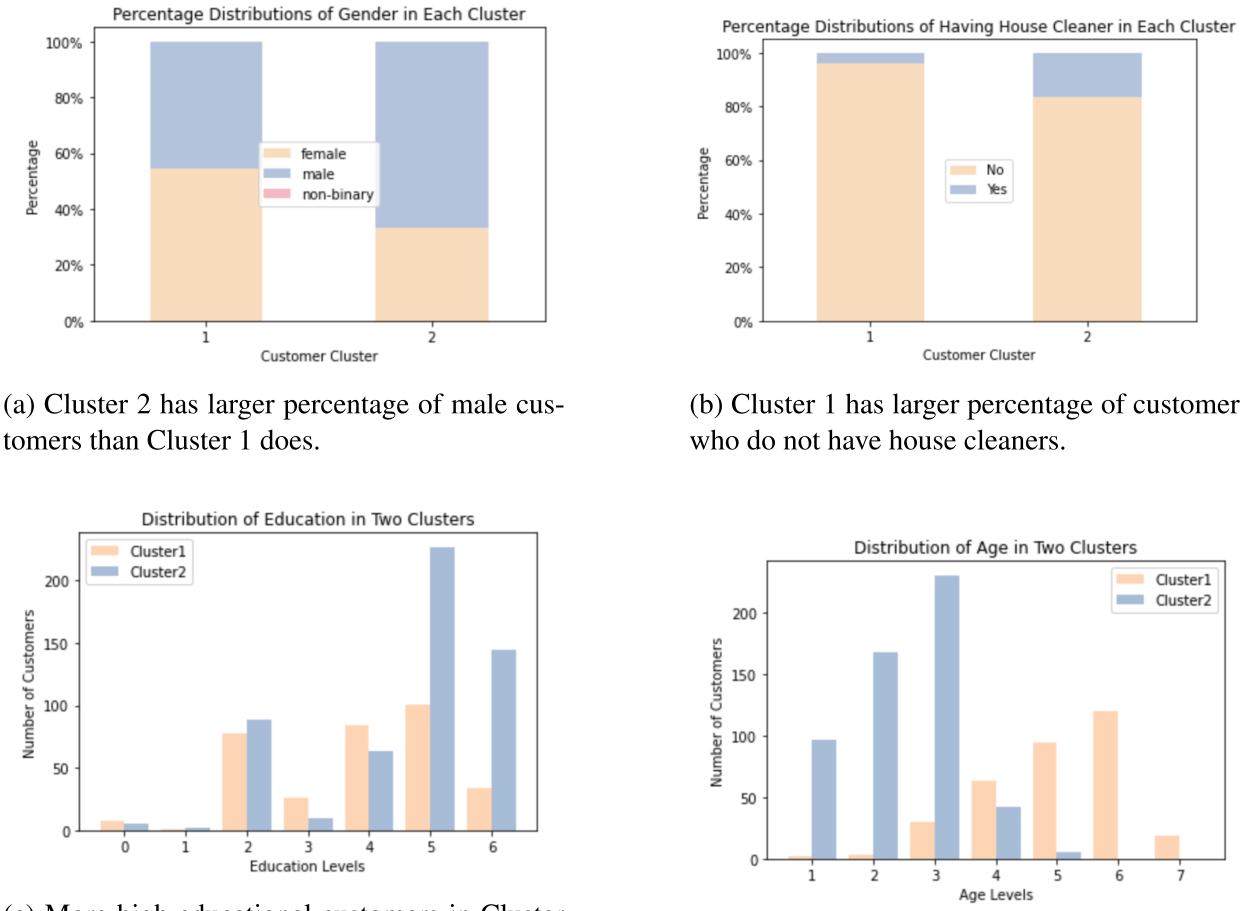


(a) Silhouette score analysis for optimal cluster count: Two clusters yields the best results

(b) TSNE visualization of k-prototype clustering results, showing the distribution of data points in two dimensions

Figure 5.7: Customer segmentation results: (a) Silhouette score analysis; (b) T-SNE plot visualization.

clusters enjoy different characteristics or make distinctive purchase decisions. This step also provides valuable insights into the distinguishing characteristics of each group. The chi-square test is a statistical method commonly used to compare two groups on categorical variables. Here, it is employed to analyze whether two customer groups differ in each of the customer characteristics on file. The results reveal that the two customer groups show significant differences in all characteristics except for the “cooking frequency” and “house type”. To illustrate, let’s consider a few examples shown in Figure 8. Firstly, cluster 2 has a higher proportion of male customers compared to cluster 1, and also a slightly larger percentage of customers who own a cleaner. Secondly, when we compare the educational distribution of the two clusters, we can see that the majority of customers in cluster 1 have education levels of 2, 4, and 5, whereas customers in cluster 2 have a slightly right-shifted peak and their education levels concentrated at 2, 5, and 6. Finally, the two clusters exhibit distinct characteristics in terms of age. Cluster 1 is composed of customers who are older than those in Cluster 2.



(a) Cluster 2 has larger percentage of male customers than Cluster 1 does.

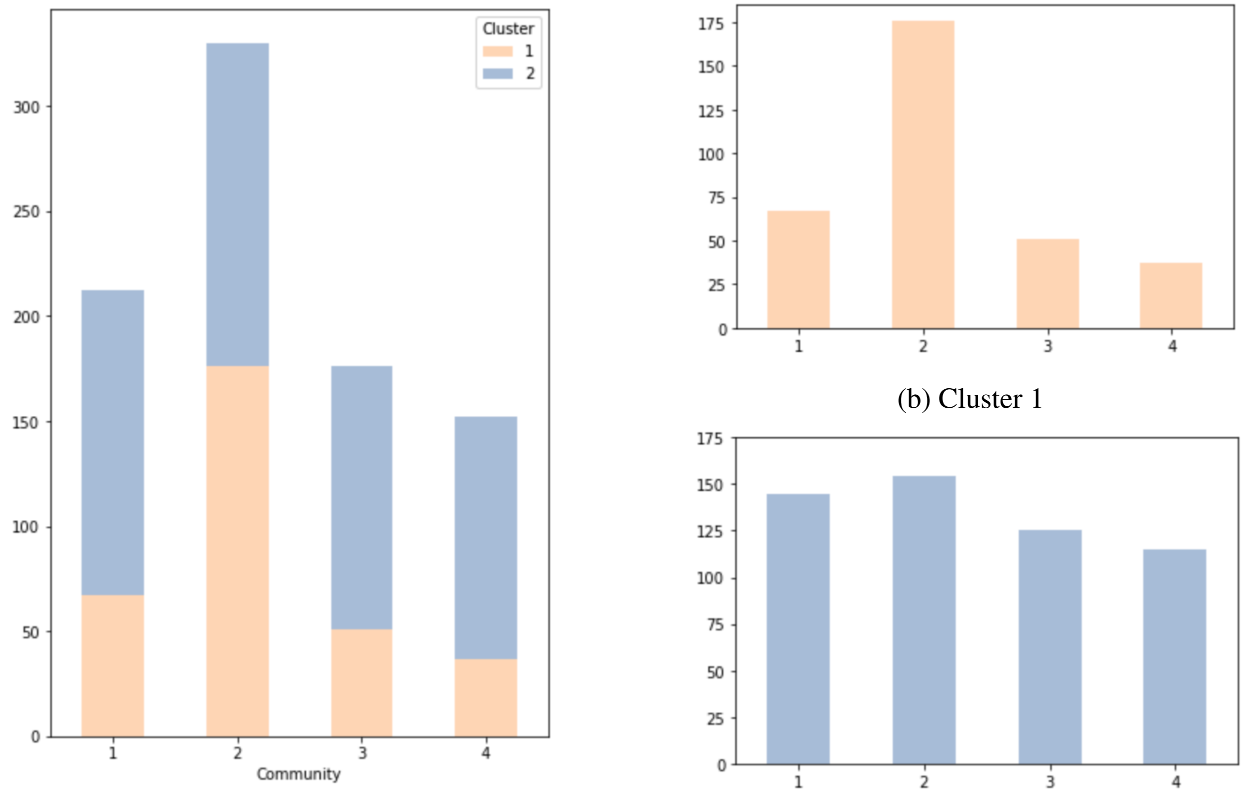
(b) Cluster 1 has larger percentage of customers who do not have house cleaners.

(c) More high educational customers in Cluster 2

(d) Cluster 2 has a younger group of customers

Figure 5.8: Comparison of Two Clusters on Selected Features

Subsequently, the study investigates whether two customer groups purchase different products. To this end, a chi-square test is conducted on customer clusters and the product communities to which their purchases belonged. The results show distinct purchasing patterns between the two clusters of customers. Specifically, over half of the customers in Cluster 2 predominantly purchased vacuum cleaners belonging to Product Community 2 (traditional and affordable products). On the other hand, while Cluster 2 customers distributed their purchases more evenly among Product Community 1 to 4, they are responsible for the majority of purchases from Product Community 4 (innovative and portable). In contrast, Cluster 1 accounted for less than a quarter of the purchases in Community 4. This comparison is reflected in part (a) figure, where cluster 2 occupies most of the bar chart for product community 4. In summary, customers of Cluster 1 exhibit a stronger preference for products from Product Community 2 while those in Cluster 2 more frequently choose products from Product Community 4. For ease of reference, we will designate Cluster 1 as the ”price-sensitive” group and Cluster 2 as the ”innovation-passionate” group based on their preferences and product characteristics.



(b) Cluster 1

(a) Entire Data

(c) Cluster 2

Figure 5.9: Distribution of Customer Purchases Across Product Communities

The tests yield important information about the distinct attributes of each segment, confirming the effectiveness of the segmentation. This emphasizes the possibility of creating personalized marketing and engagement approaches that suit the particular requirements and preferences of each customer group, ultimately enhancing targeted efforts.

### 5.3.3 Bipartite network construction, modeling, and prediction

*Bipartite network construction for different customer segments*

After removing customers who have only considered and purchased cars that are isolated in the co-consideration network, there are 871 vacuum cleaner customers that are classified into two market segments. We construct separate bipartite networks to analyze each market segment's customer preferences toward product attributes. For validation purposes, 10% of customers are reserved as testing datasets, meaning that their consideration and purchase information, which are represented as links in the constructed networks, are unknown. The customer-product bipartite networks at the consideration stage and purchase stage are plotted in Figure 5.10, where black dots are customers and colored dots are products associated with different communities. Even though customers may consider multiple products, they only purchase one from the consideration set. That is why the network for stage 1 (consideration) is much denser than that for stage 2 (choice), and the links in the choice network are conditional on the consideration stage. The size of each vacuum cleaner node reflects its level of popularity, which is determined by the frequency of its consideration or selection within the network. In both the consideration stage and the choice stage, the products from community 2 (traditional and affordable) are more prevalent among customers in cluster 1 (price sensitive).

The networks constructed for customer clusters are characterized as follows: Cluster 1 (price-sensitive) 's network comprises 331 customers who considered 286 vacuum cleaner models, with a consideration network density of 0.0031 and a choice network density of 0.0016. Network den-

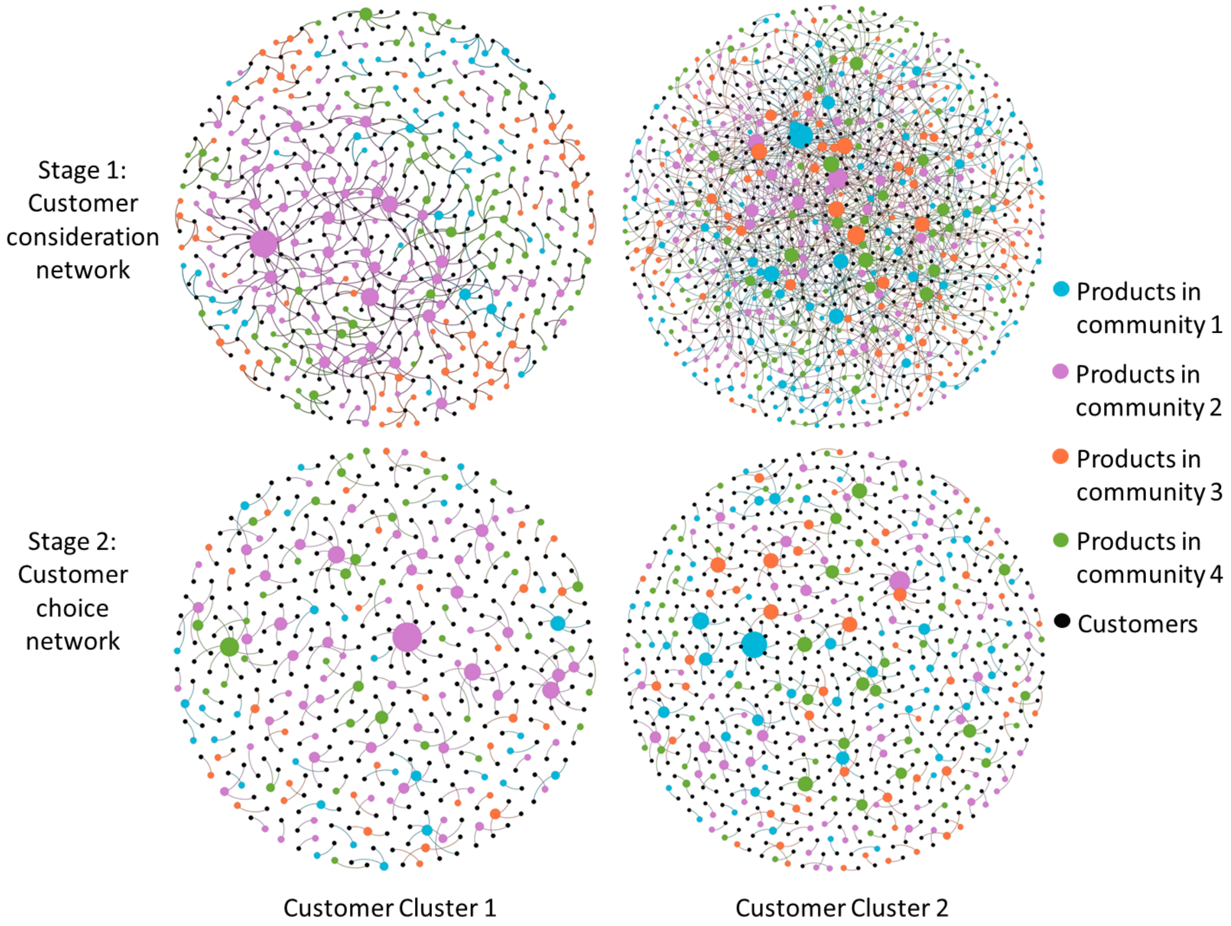


Figure 5.10: Bipartite consideration and choice networks in different customer clusters. Products in Community 2 are more prevalent in Customer Cluster 1.

sity, in this context, represents the proportion of existing connections between nodes (customers and products) relative to the total possible connections in the network. In comparison, Cluster 2 (innovation passionate) 's network includes 540 customers who evaluated 467 models, with a consideration network density of 0.0025 and a choice network density of 0.0010. The data reveals that the two networks have comparable densities, although the network for cluster 2 exhibits a larger size in comparison to the network for cluster 1.

*Model estimation and interpretation*

Prior to running the ERGM model, all numerical product attributes are normalized to a range of [0,1]. This step not only accelerates the optimization process but also simplifies the comparison of the impacts of different features in the ERGM model. Table 5.3 presents the estimated parameters (with significance levels) of the Exponential Random Graph Model (ERGM) used in the two-stage decision-making process for two distinct customer clusters. The ERGM model comprises three different levels of effects:

- **Network structure effects**, which assess the prevalence of specific network structures and test the hypothesis about common market structures studied in economics. For example, we can test whether a market is oligopoly, where a few company dominate the entire market, or competitive, where several companies vie for market share;
- **Nodal attributes**, which measure the influence of product attributes on customers' consideration and choice stages;
- **Homophily effects**, refer to the tendency of customers to consider products with similar attributes.

The **network structural effects** include three elements in our model: "edges", "market distribution", and "product shared partner". The "edges" measures the number of edges in a network and served as a baseline characteristic. Figure 5.11 provides a graphical illustration of the other two network structure effects and their interpretation in customer-product networks. The "market distribution" is measured with an endogenous structural variable of ERGM – geometrically weighted degree distribution (gwb2degree) (Hunter et al., 2008). This variable characterizes the

Table 5.3: ERGM results of two-stage modeling results for different customer clusters (market segmentation)

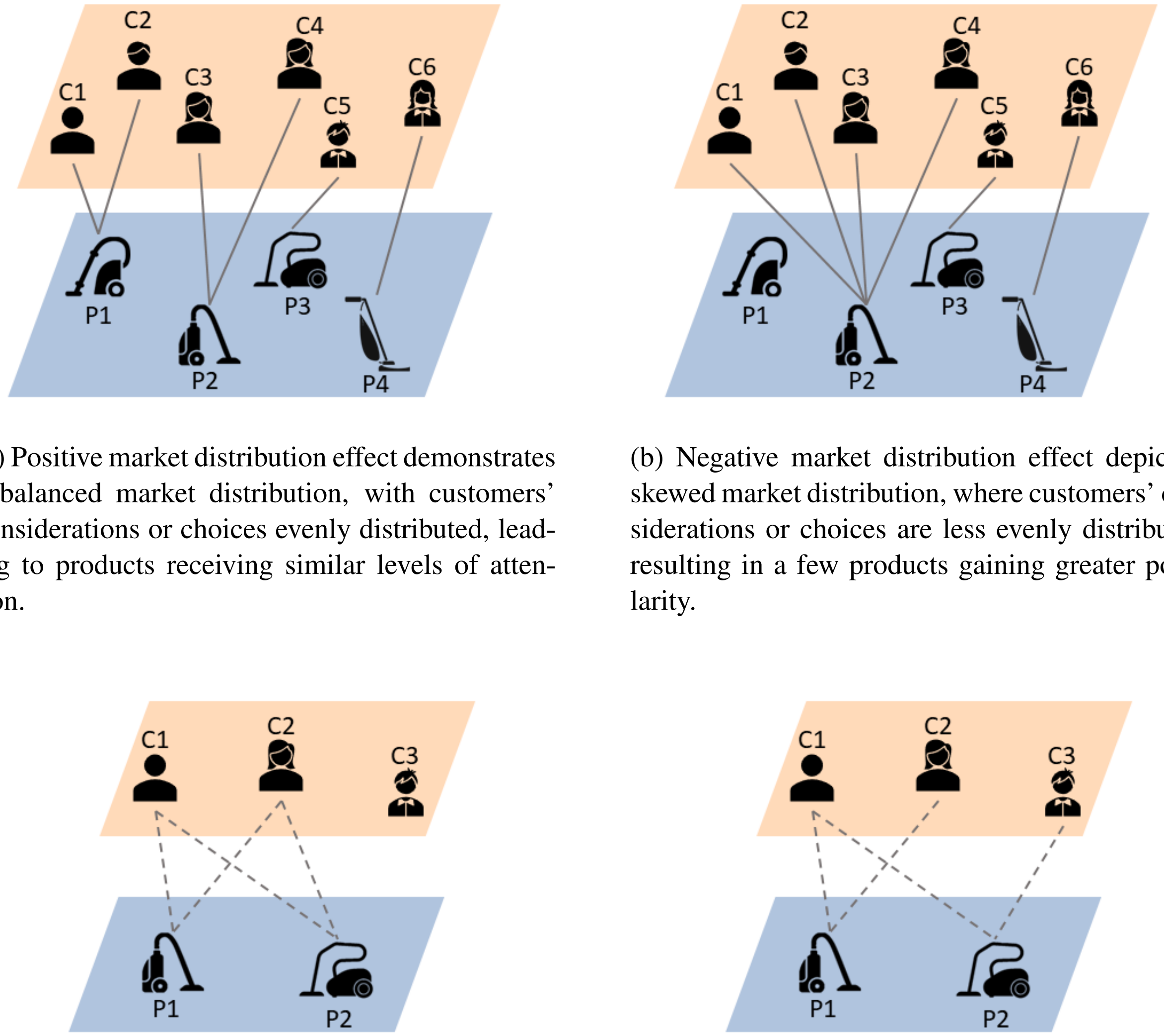
| | Cluster 1: Price sensitive | | Cluster 2: Innovation passionate | |
|---|---|---|---|---|
| Model terms | Stage 1 | Stage 2 | Stage 1 | Stage 2 |
| Edges | -6.0440*** | 1.0120. | -5.7270*** | -0.7479* |
| Market distribution | 1.1770*** | -0.7747** | -0.4156* | -0.3547* |
| Product shared partners | -0.3785*** | | -0.3745*** | |
| Upright vacuum | 0.7580*** | -0.0763 | 0.2183*** | 0.6046*** |
| Robotic vacuum | 0.3343. | -0.0824 | 0.4267*** | 0.0244 |
| Price | -0.0008. | 0.0001 | 0.0004** | 0.0002 |
| Filter: HEPA | -0.3595** | 0.2484 | -0.1533* | 0.0612 |
| Capacity | 0.0256 | -0.0363 | 0.0231*** | 0.0273* |
| Bagless | 0.5097* | -0.1515 | 0.4910*** | 0.3286 |
| Suction power | 0.1264. | -0.0919 | 0.1620*** | -0.0771 |
| Surface match | 0.2108** | | 0.2623*** | |
| Model fit: AIC | 6997 | 1427 | 15268 | 2609 |
| Model fit: BIC | 7109 | 1511 | 15392 | 2712 |

Note. .$p < .1$; *$p < .05$; **$p < .01$; ***$p < 0.001$

distribution of the degrees among the vacuum cleaner models. A positive coefficient for market distribution indicates a more even distribution among degrees (i.e., most vacuum cleaners have similar sales, as depicted in Figure 5.11a), while a negative coefficient implies a skewed distribution (i.e., a few vacuum cleaners have significantly higher sales than others as illustrated in Figure 5.11b). Notably, for customer cluster 1 at the consideration stage, the market distribution exhibits a positive and significant effect, implying that most vacuum cleaners have comparable chances of being considered by customers. In contrast, for customer cluster 2, the market is more skewed in the consideration stage. This trend is further evidenced by the product degree distribution plot for the consideration stage shown in Figure 5.12. In this plot, product nodes within Cluster 2 exhibit a long-tail effect, indicating that a small number of products receive significantly more attention compared to the rest. However, in the choice stage, both networks in Cluster 1 and Cluster 2 are skewed, indicating that final sales are more concentrated towards a few products in the final choice.

The “product shared partner” is measured by geometrically weighted dyadwise shared partner distribution (gwb2dsp) (Hunter et al., 2008), which is specifically relevant to consideration networks. This metric helps us understand how pairs of product nodes that have shared connections to some customers tend to form connections with other customer nodes. In other words, it helps us see how products that are popular with the same customers might be related to each other. Intuitively, if two products are co-considered by a customer, they are more likely to be co-considered by additional customers, suggesting that customers tend to consider the same products. However, the model results indicate negative coefficients for product shared partners, revealing that products are often co-considered by a few people (as illustrated in Figure 5.11d). This observation can be attributed to the limited size of the customer sample in the survey data, which results in a lack of shared considerations among customers. Additionally, the diverse nature of the vacuum cleaner market may contribute to this phenomenon, as a wide variety of available products prevents customer considerations from concentrating on just a few options.

The estimation of **nodal attributes** highlights the significance of vacuum cleaner product attributes in the customer consideration and choice stages. A positive effect for nodal attributes indicates that a product with these attributes is more likely to be considered or chosen. In the consideration stage, both upright and robotic vacuum cleaners are popular for both customer clusters. However, innovation-passionate customers (cluster 2) demonstrate a much stronger preference for robotic vacuum cleaners, while price-sensitive customers (cluster 1) favor upright vacuum cleaners. In terms of “price”, customer cluster 1 is more price-sensitive, as the price has a negative effect in their consideration stage, while customer cluster 2 demonstrates lower price sensitivity. Regarding engineering attributes, it appears that vacuum cleaners with “HEPA filters” (more efficient filters) do not significantly increase customers’ interest in the consideration stage. Although both customer clusters favor “bagless” and strong “suction power” products, customer cluster 2



(a) Positive market distribution effect demonstrates a balanced market distribution, with customers' considerations or choices evenly distributed, leading to products receiving similar levels of attention.

(b) Negative market distribution effect depicts a skewed market distribution, where customers' considerations or choices are less evenly distributed, resulting in a few products gaining greater popularity.

(c) Positive product shared partner effect highlights the tendency for pairs of products (P1 and P2) that are co-considered by one customer to be more likely co-considered by other customers as well.

(d) Negative product shared partner effect illustrates the scenario where pairs of products (P1 and P2) that are co-considered by one customer become less likely to be co-considered by other customers.

Figure 5.11: Illustration of network structural effects in customer-product networks with distinct layers for customers and products.

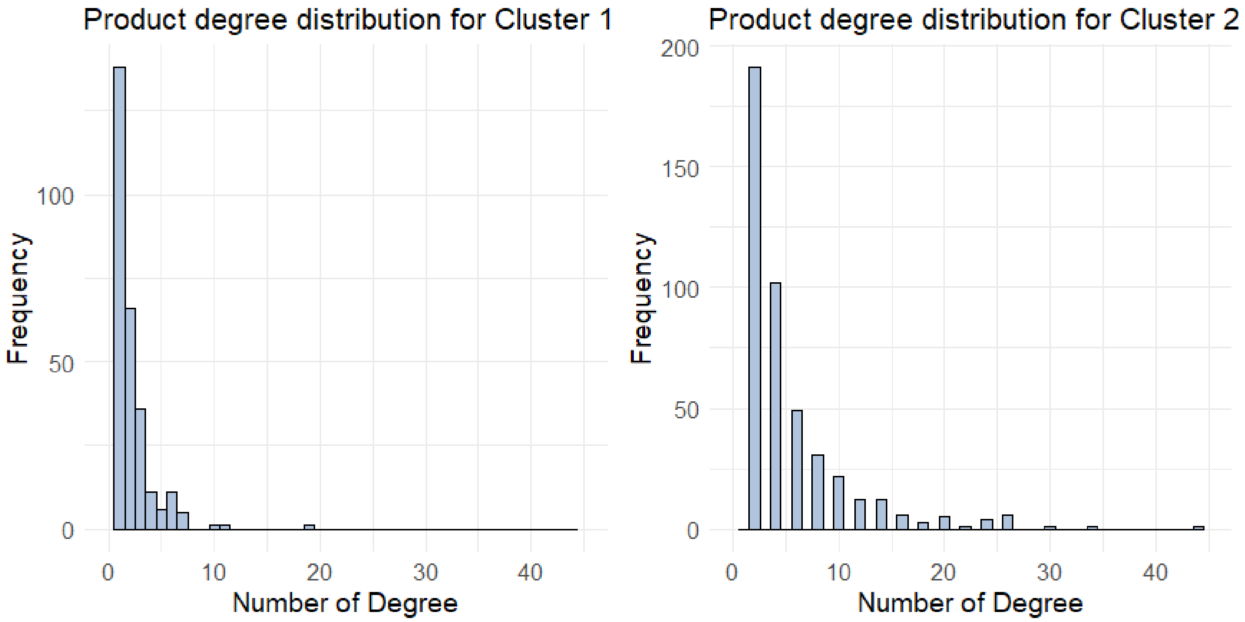


Figure 5.12: Degree distribution of products in consideration networks for customer clusters. Customer cluster 2 exhibits a long-tail effect, with a few highly popular products, while products in customer cluster 1 have more similar degrees, reflecting comparable popularity.

exhibits a stronger preference compared to cluster 1. Furthermore, they also show a keen interest in vacuum cleaners with larger "capacities".

We also observe that in the choice stage, customer cluster 2 finds upright vacuum cleaners more appealing. This implies that despite a strong interest in innovation-passionate customers to consider robotic vacuum cleaners, they often end up selecting upright vacuum cleaners from their consideration set. This preference is supported by the fact that among all survey respondents, 47.03% of them selected upright vacuum cleaners, while the remaining 52.97% chose from robotic, stick, handheld, and canister vacuum cleaners. Meanwhile, product features are less influential in customer decision-making during the choice stage compared to the consideration stage. Only "capacity" has marginally significant effects on customer cluster 2. This may be because the model cannot take into account other important factors, such as customer ratings and online recommendations, which can influence the choice stage. Additionally, most product features have already been evaluated during the earlier consideration stage.

The network models effectively capture **homophily effects**, which are analyzed during the consideration stage to determine if product similarity influences customers to consider them together. By incorporating the model term “surface match”, we observe significant and positive impacts of the feature, indicating that customers are more likely to consider products with the same ”surface recommendation” (e.g., hardwood, carpet).

By comparing the key factors in the customer decision-making process across multiple stages, we can discern the different preferences between the two distinct customer clusters. Customers in the ”innovation-passionate” cluster place greater emphasis on product features and are willing to pay more for superior products. In contrast, customers in the ”price-sensitive” cluster exhibit a stronger preference for upright vacuum cleaners with more traditional designs. Furthermore, product features exert a more significant influence during the consideration stage, while their impact diminishes during the choice stage.

We also run the baseline network model that does not involve market segmentation. The estimated parameters in each stage are recorded in Table 5.4. The single bipartite network uses data that composes 871 customers and 528 products. We notice that the obtained coefficient of market distribution (-1.0540) in the single network at the consideration stage is negative, meaning that the market distribution is more skewed, and some products are more frequently considered than the rest of the products in the market. On the contrary, with market-segment-based network modeling, the products are more evenly considered by customers in customer cluster 1. This finding suggests that, when viewed holistically, the vacuum cleaner market is dominated by a few key players. The market-segmentation-based models, on the other hand, enable us stratify two distinct sub-markets - one that is highly competitive and another that remains monopolistic. For preference towards vacuum cleaner type, the single model shows that both upright and robotic vacuum cleaners are preferred compared to other types of vacuum cleaners but it can't differentiate the specific product attributes that are influential for each type. The market-segment-based models overcome this difficulty and we can observe the differences in preferences for upright and robotic vacuum cleaners for different customer clusters in Table 5.3. This observation also applies to the investigated product features. Although the baseline model exhibits similar trends for feature importance as the market-segmented model, it fails to differentiate varying preferences for each product feature among different customer clusters.

Table 5.4: ERGM results of two-stage modeling results for a single network model (without market segmentation)

| Model terms | Stage 1 (consideration) | Stage 2 (choice) |
|---|---|---|
| Edges | -5.8230*** | -0.2863 |
| Market distribution | -1.0540*** | -0.5529*** |
| Product shared partner | -0.3418*** | |
| Upright vacuum | 0.3173*** | 0.4351*** |
| Robotic vacuum | 0.3340** | -0.0622 |
| Price | 0.0002. | 0.0001 |
| Filter: HEPA | -0.2112** | 0.0892 |
| Capacity | 0.0169*** | 0.0177 |
| Bagless | 0.4079*** | 0.1977 |
| Suction power | 0.1341*** | -0.0941 |
| Surface match | 0.2188*** | |
| model fit: AIC | 23288 | 4023 |
| model fit: BIC | 23420 | 4132 |

Note. $.p < .1$; $*p < .05$; $**p < .01$; $***p < 0.001$

*Model prediction and validation*

While we have reported the Akaike information criterion (AIC) and Bayesian information criterion (BIC) values for the model fit measurement in Table 5.3 and 5.4, a direct comparison between network models with and without market segmentation is not feasible. This is because these metrics are only comparable when the model structure and input data are identical. They are primarily useful for evaluating the goodness of fit when incorporating different model terms within a single model that uses the same data.

Therefore, to validate our market-segmentation-based model's performance compared with the benchmark (a single network), we conducted a prediction analysis for both the consideration and choice stages on the testing customers. These customers comprise 10% of the total and are considered to have missing edges towards products. This comparative analysis allows us to assess the effectiveness of our market-segment-based network modeling in relation to the baseline single network approach, providing valuable insights into the accuracy and predictive capabilities of the two methods at each stage of the decision-making process.

Upon estimating the effects for the specified model terms in the ERGM model, we simulate the entire network based on the ERGM formula in Equation 2.5, using the Gibbs sampling technique, which is embedded in the Statnet package for ERGM. Due to the stochastic nature of the simulation, we generate 500 networks and calculate the average probability of link existence. This approach allows us to predict the probabilities of missing edges while preserving the training customers' links as observed in the original network.

In the consideration stage, we identify the products most likely to be considered by each customer based on the probabilities of the simulated links. We rank the potential products according to their likelihood of being considered and identify the Top N. These products are then compared to the actual products considered, as reported by the testing customers. We calculate a hit rate by determining the percentage of actual products covered by the top N product sets. This hit rate allows us to evaluate the overall accuracy of the model. Figure 5.13 illustrates the hit rate of product consideration prediction for different models, with both random prediction and network without segmentation serving as benchmark models. The area under the curve has also been calculated as a quantitative indicator of the model's overall performance; an area equal to 1 represents a perfect and accurate prediction. As shown in the figure, the network with segmentation improves the overall predicted hit rate by 21.2%.

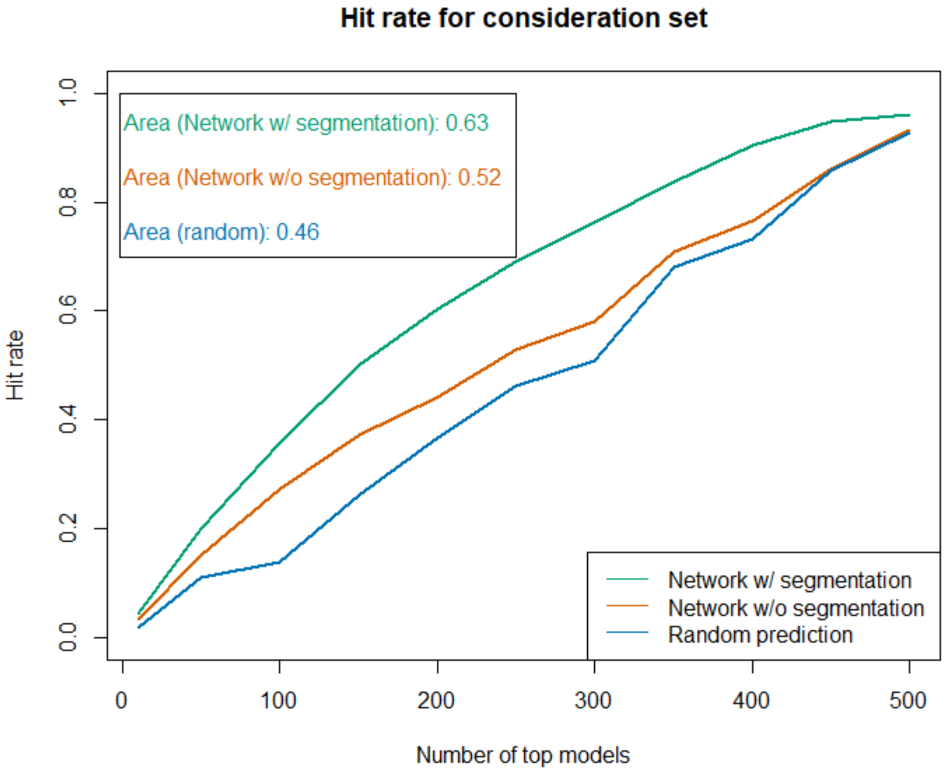


Figure 5.13: Comparison of hit rate performance for three different models: (1) random selection, (2) single network without market segmentation, and (3) market-segmented model

In the choice stage, we also simulate the network, but this time we restrict each customer to form a purchase link only within their actual consideration set. In other words, we predict customers' choices conditional on their consideration sets. By calculating the probability of purchase link existence using 500 simulated networks, we identify the product with the highest probability as the customer's choice within their consideration set. In the original survey data, there are respondents who reported only one product in their consideration set, and that product was also the one they ultimately purchased. As such, predicting their choices based on their consideration set is not meaningful in these cases. Therefore, when calculating prediction accuracy, we remove customers who only consider one product. The remaining customers have two to seven products in their consideration sets. As a baseline, random guessing predicts 36.67% of customer choices correctly, and the single network model without segmentation yields 50% correct predictions. The market segmentation model further improves prediction accuracy to 56.32%.

In summary, network-based models have demonstrated their predictive capabilities in forecasting customer considerations and choices, given the ERGM model and the inherent network structures. Additionally, the predictive power of the market-segmentation-based model surpasses that of the single network model without segmentation, indicating that market segments effectively capture the heterogeneity in customer preferences. There are still limitations to the predictive power of network-based models, which are addressed in the discussion section of our study.

## 5.4 Discussion

The prior section highlights the efficacy of our market-segmentation-based network modeling approach. Next, we delve into its implications on market understanding and product design, while also addressing limitations and suggesting potential improvements.

### 5.4.1 Implication on market understanding and product design

The market-segmentation-based model proposed in this study enhances the applicability of network-based methods in customer preference modeling, particularly for diverse and mixed product markets. This approach stratified the market into finer layers and allow market practitioners to locate and better serve their target customers. When the product market is treated as a single entity, only overarching patterns are discernible. However, breaking down the market into distinct segments uncovers unique competitive dynamics of each sub-market, providing a more profound understanding of the market's overall structure. This empowers market practitioners to formulate strategies tailored to specific sub-markets. For product designers, the ability of the customer preference model to accommodate heterogeneity is crucial. It facilitates the development of products tailored to a range of customer preferences, aligning design with market positioning.

This approach not only caters to diverse customer needs but also offers a more comprehensive understanding of the two-stage decision-making process. By examining various factors that play critical roles at different stages, marketers and product designers can develop tailored strategies accordingly. For instance, if products are often considered by customers but rarely chosen, it is essential to identify the factors that influence customers' choices during the evaluation of their consideration set. By addressing these factors, businesses can better cater to customers and improve their products' chances of being chosen.

### 5.4.2 Limitations

Building on the implication of our research, it is essential to acknowledge some limitations that may impact the findings of this study. First, our survey data is limited in size, which may impact the robustness of our results. A larger dataset that captures a more diverse range of customer preferences and product markets would provide a more comprehensive understanding of the relationships identified in our study. Second, while the network-based model demonstrates effectiveness in capturing the interdependence of the customer decision-making process and outperforms traditional choice models, its predictive capability has inherent limitations. The linear nature of the model restricts its ability to make highly accurate predictions, particularly when compared to more advanced machine learning models that can accommodate complex nonlinear relationships. Despite these limitations, our research offers a valuable foundation for future studies to build upon. Researchers may consider expanding the dataset size and exploring alternative modeling approaches to enhance the predictive power of network-based customer preference models.

## 5.5 Conclusion

In this research, we propose a novel approach for identifying heterogeneous customer preferences in the two-stage consideration-then-choice decision-making process. More specifically, we first used the Joint Correspondence Analysis to analyze the relationship between product association communities and customer attributes (demographic, usage context, and personal viewpoints) in a low-dimensional space. We pinpoint important customer attributes through the process and segment the market into clusters based on these attributes. For each identified market segment, we construct a bipartite customer-product network that denotes the customer-product relations in the customers' consideration and choice stages respectively, given that the choice stage network is conditional on the consideration stage to mimic customers' decision-making process. Finally, by adapting the Exponential Random Graph Model, we investigate how various factors influence customer decision-making processes and how they differ between distinct customer groups.

Our analysis of real customer survey data for the vacuum cleaner market indicates that product attributes play more important roles in the consideration stage compared to the choice stage. Additionally, the same product attributes could have varying effects on different market segments (innovation-passionate customers versus price-sensitive customers in our case study). We further validate the market-segmentation-based network models by comparing their predictive power to a benchmark model. The results show that network-based models with market segmentation not only offer a more practical interpretation of customer preferences by reflecting diverse customer tastes compared to a single network model, but also provide more accurate estimations of customers' considerations and choices.

To the best of the authors' knowledge, this study represents the first exploration in network-based customer preference modeling that focuses on customer segmentation to gain valuable insights into the intricate and diverse nature of customer preferences. The contributions of the research methodology can be outlined as follows: First, traditional study segments a market by using either specific aspects of customer attributes or all available customer attributes without examining their relevance and association with preferences. In our study, we propose a data-driven approach that selects meaningful customer attributes reflecting demographic attributes, usage contexts, and personal viewpoints to identify market segmentation. Secondly, our model introduces a network-based approach across various stages of the customer's decision-making process within a diverse and mixed product market. This represents the first application of network-based customer preference modeling beyond its initial test dataset (i.e., the car market). Enabled by market segmentation, our model can capture distinct market patterns and heterogeneous customer preferences within two sub-networks. Importantly, it also uncovers unique factors influencing customers during the consideration and choice stages of their decision-making process. Furthermore, this study presents a new validation method for network-based models based on network predictions. This method preserves a portion of network links as missing edges and predicts the network by simulating new networks. This validation approach enables the comparison of the performance of network-based models with different data and structures. In our study, we demonstrate the effectiveness of our proposed model by comparing it with a benchmark model using this validation method. In summary, our method presents a systematic market-segmentation-based network approach for investigating customer preferences in their consideration-then-choice decision-making behavior. This research lays the groundwork for future studies exploring more comprehensive network-based methods for examining customer preferences across different product markets.

In the previous and current chapters, we employed ERGM-based methods to model customer preferences. However, it is crucial to highlight that the ERGM model exhibited limited prediction accuracy, as evidenced by the results presented. This observation necessitates the exploration of alternative approaches, particularly deep learning models, to enhance the accuracy of the models utilized. Consequently, the forthcoming chapter will delve into the utilization of deep learning models as a means to address this limitation, improve prediction accuracy, and provide valuable insights into the complex and heterogeneous nature of customer preferences.

# CHAPTER 6
# GRAPH NEURAL NETWORK BASED METHODS IN LINK PREDICTION

## 6.1 Introduction

Complex engineering systems encompass a multitude of stakeholders and entities that are interconnected through intricate relationships. The comprehension and accurate prediction of these relationships are paramount for the effective study and manipulation of these systems. An example of a complex engineering system is the car market, where there are many interactions between stakeholders. The success of a new car depends not only on its engineering performance but also on its competitiveness relative to similar cars and factors such as perceived market position. Customers from different geographies may prefer different types of vehicles. A design intervention in the car market, either by introducing changes in existing cars or launching a new car design, may encourage customers to change their driving behavior. When these changes happen, a manufacturer needs to understand which products their car models will compete in the new situation and what they can do to improve their market position. It is also important to consider the complex relationship among customers, such as the social network between customers and the complex relationships among products. Network analysis is a crucial method for statistical analysis of complex systems in many scientific, social, and engineering domains (Holling, 2001; M. E. Newman, 2003; Simon, 1977; Wasserman & Faust, 1994).

Researchers have employed exponential random graph models (ERGMs) as a statistical inference framework to interpret complex customer-product relations. ERGMs have been employed in literature to study customers' consideration behaviors (Sha, Saeger, et al., 2017; M. Wang, Chen, Huang, et al., 2016). These studies illustrated the benefits of using the network-based preference model for predicting the outcome of design decisions. However, ERGMs have a few limitations. First, they are typically appropriate for small to medium-sized networks with a few attributes. For large datasets, the MCMC approach to estimate ERGM parameters does not converge (as the work in chapter 3). This leads to an important limitation for product manufacturers, who now want to make the most of massive datasets but still want statistical models to help them understand what is happening inside these models. In addition, previously published research shows that future market forecasts based on ERGMs are not sufficiently accurate at capturing the true network (Sha, Huang, Fu, et al., 2018). Poor forecasts can affect the manufacturer's market position as inaccurate predictions of the market competition can lead a manufacturer to wrongly estimate their future market position when they introduce a new car or change a feature in an existing car. If manufacturers rely on poor predictions to introduce design changes, the result will also affect the customers, as the new choices present in the market may lack what they desire. If car manufacturers have a method to predict product competition for future years or customers' choices accurately, they can also use these predictions to identify competitors and incorporate them in designing their strategy for product placement, marketing, or redesign of the car. Manufacturers can also estimate how their market position may change when competitors introduce changes in existing attributes. This chapter presents such an approach by modeling customer-product networks using deep learning approaches, which does not face the issues highlighted above.

Against this backdrop, Graph Neural Networks (GNNs) have emerged as a promising solution due to their ability to model both discrete and continuous representations and their broad expressive power (Zhou et al., 2018). Their versatility is reflected in their successful deployment across a range of domains, including drug discovery, image classification, natural language processing, and social network analysis (Stokes et al., 2020; Z. Wu et al., 2020). Moreover, GNNs offer clear advantages over traditional unstructured machine learning methods due to their support for interpretability, causality, and inductive generalization. They have been deployed by major corporations such as Uber Eats and Alibaba for tasks such as food and product recommendation, respectively (Jain et al., 2019; J. Wang et al., 2018). Even within the field of Engineering Design, albeit less commonly, GNNs have been utilized for tasks such as product tolerance design, machining feature recognition, and understanding mechanical device function (W. Cao et al., 2020; Li et al., 2021; J. Wang et al., 2020). In this chapter, we aspire to apply GNNs to model customer-product networks, offering a viable alternative to the limitations of ERGMs and providing accurate predictions of product competition and customer choices. This enables manufacturers to formulate effective strategies for product placement, marketing, or vehicle redesign and to anticipate potential changes in their market position due to modifications in existing product attributes.

The primary emphasis of this chapter is the application of GNNs for link prediction in two distinct, yet interconnected, scenarios. The first scenario explores a unidimensional network with a focus on product competition networks. The second scenario, on the other hand, investigates a bipartite network that includes customer nodes and product nodes. This investigation aims to untangle the complexities inherent in a two-stage customer decision-making process.

## 6.2 Link prediction of product competition network

### 6.2.1 Methodology

In this work, We establish a product co-consideration network to model product competition behavior and use a GNN approach to predict future product competition. The methodology of the training and prediction process for the link existence is shown in Figure 6.1.

The methodology comprises five main components as follows:

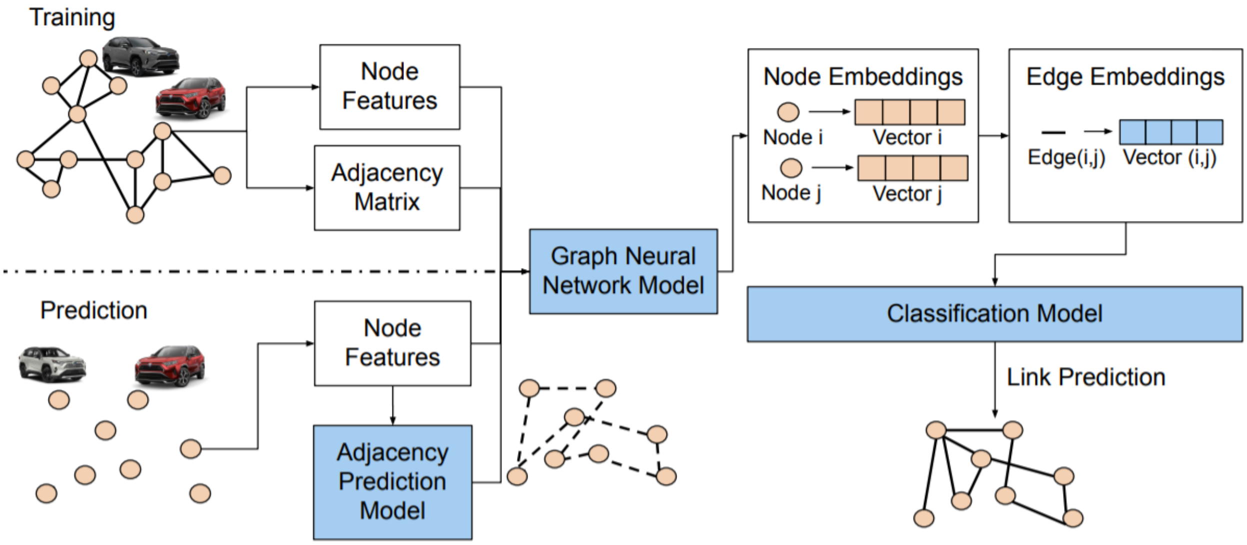


Figure 6.1: The methodology of predicting the link existence in a car competition network using graph neural network model

1. **Representing products and their relationships as a graph**: this step involves the data processing and transformation to construct a network with products as nodes and their relationships as links.

2. **Training the GNN to learn the graph structure**: this step finds a low-dimensional embedding of nodes and edges in the contracted graph.

3. **Training classification models to make predictions**: this step takes the graph embeddings as input to train a classification model on link existence.

4. **Creating an adjacency prediction model to augment the GNN for unseen data**: for validation, the model is tested on the held-out network and unseen network. A proposed adjacency prediction model is applied in the unseen network prediction.

5. **Interpreting the importance of design attributes**: based on the model, this step investi-

gates the importance of the features and provides useful insight for the engineering design.

*Inductive representation learning on networks*

Many GNN models can learn functions trained on a graph and generate the embeddings for a node, which sample and aggregate feature and topological information from a node's neighborhood. However, engineering applications require methods that can make predictions about completely new nodes too. This need inspired us to employ GraphSAGE— a representation learning technique for dynamic graphs, which learns aggregator functions that can calculate new node embedding based on the features and neighborhood of a node.

As illustrated in Figure 6.2, GraphSAGE learns node embeddings for attributed graphs (where nodes have features or attributes) through aggregating neighboring node attributes. The aggregation parameters are learned by the ML model by encouraging node pairs co-occurring in short random walks to have similar representations.

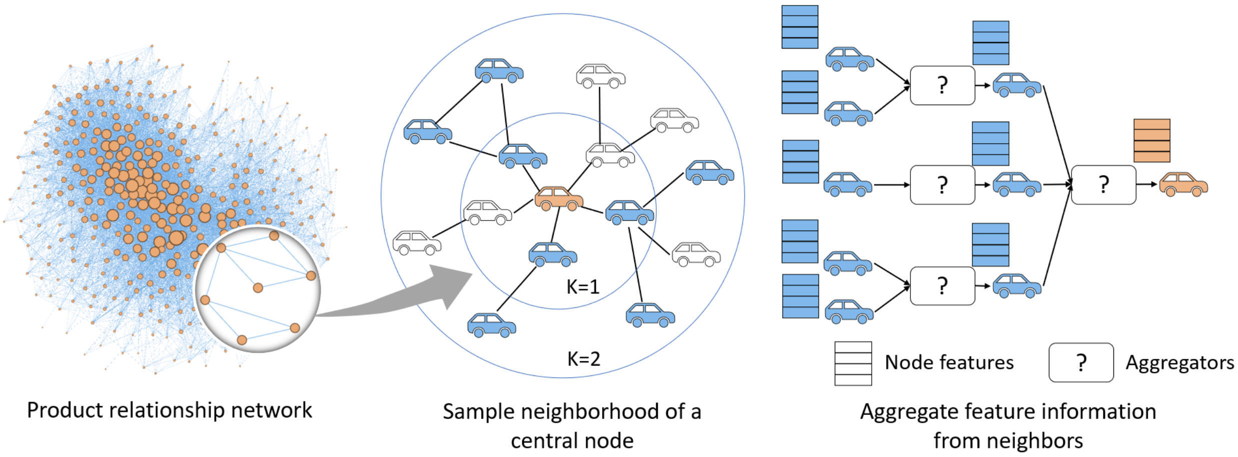

K=1
K=2
?
Node features
Aggregators
Product relationship network
Sample neighborhood of a central node
Aggregate feature information from neighbors

Figure 6.2: Illustration of sampling and aggregation in GraphSAGE method. A sample of neighboring nodes contributes to the embedding of the central node.

The detailed algorithm of GraphSAGE from (Hamilton et al., 2017) is shown in Algorithm

1. In GraphSAGE, it is assumed that every node can be defined by its neighbors, which means that the embedding for a node can be calculated by some combination of the embedding vectors of its neighbors. At the beginning of the training, every node's embedding is set equal to its feature vectors. The algorithm follows two main steps — aggregate and update (Step 4 and 5 in Algorithm 1). The aggregate step uses any differentiable function to aggregate the embedding of neighbors to find the embedding of the target node. A typical example of the aggregate step can be simple averaging of neighbors. The update step uses a differentiable function to combine the new aggregated representation for the target node with its previous representation. The $K$ parameter tells the algorithms how many neighborhoods or hops to use to compute the representation for the target node. The aggregation can occur for first neighbors ($K = 1$) or from neighbors that are further away ($K \geq 1$). However, if too many neighbors at different depths are used, that may dilute the effect of a local neighborhood. On the other hand, if only the first neighbors are considered, the method will be equivalent to using a simple neural network. Interested readers are encouraged to read (Hamilton et al., 2017) for details of the algorithm.

**Node embeddings** To train a GraphSAGE model, the inputs are the product attributes (i.e., node features) and the network structure (i.e., adjacency matrix) of the product co-consideration network. Then for each node, the GNN models can encode nodes into lower-dimensional space in the node embedding stage. For example, as illustrated in Fig.6.1, nodes $i$ and $j$ can be represented by vectors $i$ and $j$, which carry the information of node $i$'s and $j$'s features and local neighborhoods, respectively.

**Edge embeddings** Using a GNN-trained embedding for nodes, one can also learn the representation for all possible links (edges) in the network. Learning link representations is done by aggregating every possible pair of node embeddings. We use the dot product of vectors $i$ and $j$

**Algorithm 1:** Embedding generation (i.e., forward propagation) algorithm from (Hamilton et al., 2017)

```
Input  : Graph $G(\mathcal{V}, \mathcal{E})$; input features $\mathbf{x}_v, \forall v \in \mathcal{V}$; depth $K$; weight matrices
         $\mathbf{W}^k, \forall k \in \{1, ..., K\}$; non-linearity $\sigma$; differentiable aggregator functions
         $\text{AGGREGATE}_k, \forall k \in \{1, ..., K\}$; neighborhood function $\mathcal{N} : v \rightarrow 2^{\mathcal{V}}$
Output: Vector representations $\mathbf{z}_v$ for all $v \in \mathcal{V}$

$\mathbf{h}_v^0 \leftarrow \mathbf{x}_v, \forall v \in \mathcal{V}$ ;
for $k = 1...K$ do
    for $v \in \mathcal{V}$ do
        $\mathbf{h}_{\mathcal{N}(v)}^k \leftarrow \text{AGGREGATE}_k(\{\mathbf{h}_u^{k-1}, \forall u \in \mathcal{N}(v)\})$;
        $\mathbf{h}_v^k \leftarrow \sigma\left(\mathbf{W}^k \cdot \text{CONCAT}(\mathbf{h}_v^{k-1}, \mathbf{h}_{\mathcal{N}(v)}^k)\right)$
    end
    $\mathbf{h}_v^k \leftarrow \mathbf{h}_v^k / \|\mathbf{h}_v^k\|_2, \forall v \in \mathcal{V}$
end
$\mathbf{z}_v \leftarrow \mathbf{h}_v^K, \forall v \in \mathcal{V}$
```

to find the edge embeddings. Note that other symmetric operations such as averaging can also aggregate node embeddings to give an edge embedding. Our experiments found that the dot product gave slightly better results than the averaging operator (same F-1 score and 0.07 higher AUC score), which led us to select the dot product as the aggregation method in this study. Once we learn the edge embeddings, they can be used as an input to any ML model, which can be trained to predict whether an edge exists or not, which is discussed next.

*Classification model for link prediction*

The link prediction problem can be posed as a binary classification problem, where the goal is to predict whether a link candidate exists in the network (Class 1 or a positive edge) or does not exist (Class 0 or a negative edge). During the GNN model training, we can also train a downstream classification model to predict link existence, given the edge embedding as an input.

For each pair of nodes, the classification model takes the edge embeddings as input and whether the link exists or not as labels. Any classification model (such as logistic regression, k-nearest neighbors, and naive Bayes classifiers) can be integrated with the GNN model to predict the link existence. We used a multilayer perceptron (MLP) model for this work. Note that the GNN model and the MLP based classification model are trained simultaneously for the supervised learning task in the training process. To avoid imbalanced training of the classification model for networks with very few edges, we balance the two classes by sub-sampling the negative edges (an edge that does not exist in the training data).

*Permutation-based feature importance*

Besides forecasting future market competition in the engineering design domain, it is important to understand the dominant features in product competition. Therefore, we investigated the importance of different design attributes in the GNN method using "Permutation feature importance" (Molnar et al., 2020).

We used the method outlined in (Molnar et al., 2020) to measure the importance of a feature by calculating how much a model's prediction error increases on average when a particular feature is permuted randomly. A feature was considered "important" if shuffling its values significantly increased the model error. This implied that the model relied on this feature to make accurate predictions, as measured by less prediction error. A feature was considered "unimportant" if shuffling its values left the model error unchanged. This implied that the model ignored the feature for the prediction and was not dependent on it to make good predictions. The outline of the permutation importance algorithm is described in section 8.1 in (Molnar et al., 2020).

Some other methods to calculate feature importance suggest removing features, retraining the model, and then comparing the model error. In contrast, permutation feature importance does not require retraining the model. Since the retraining of an ML model can take a long time, only permuting a feature can save time and inform us of the importance of features for that particular model. This technique is independent of what ML model is used and generally, several different permutations are used to estimate the metric. One also needs to define what metric (such as the AUC value for a classification model) they are using to calculate the change in performance. This metric does not reflect the intrinsic predictive value of a feature by itself. Instead, it shows how important the feature is for a particular model.

It is noteworthy that the permutation methods on feature importance can be applied to either training data or test data. Applying it to training data will help understand how much the model relies on each feature for making predictions (training data). Applying it to test data will help understand how much the feature contributes to the performance of the trained ML model on unseen test data. Our analysis uses it for the training data as the feature importance found using test data can change if the model is tested on different test sets.

### 6.2.2 Case study

In this section, we demonstrate the use of the GNN approach to study the Chinese car market. We used car survey data provided by the Ford Motor Company as a test example. We show that by training a GraphSAGE model, we can predict the future market competition even though cars in the future may have new attributes such as increased engine size or new products may be introduced. We also show how statistical methods can be employed to calculate the importance of each attribute for the relationship prediction task.

The dataset used in this study contains 2013 and 2014 car customer survey data in the China market. In the survey, more than 40,000 respondents each year specified which cars they purchased and which cars they considered before making their final car purchase decision. Each customer indicated at least one and up to three cars which they considered. The dataset also contains attributes for each car (e.g., price, power, brand origin, and fuel consumption) and many attributes for each customer (e.g., gender, age).

*Link prediction*

This section explores various facets of link prediction utilizing the GraphSAGE algorithm, and also compares its prediction accuracy against traditional ERGM methods.

**Predicting missing links in the same year** In this part, we test our method for predicting held-out links from a network of cars from the 2013 data. We split the network into two parts to train the model by sampling a subset of links– the training graph and the test graph. Both the graphs contain the same nodes and do not contain any isolated nodes. For the training graph, an equal number of positive and negative edges were sampled to ensure that the model is trained on a balanced dataset. The test graph was used for evaluating the model's performance on held-out data.

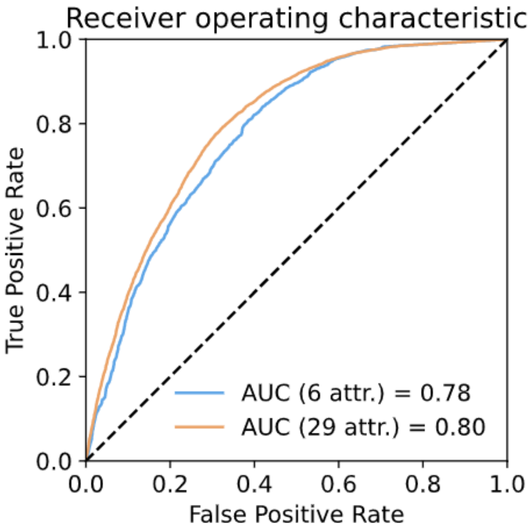


Figure 6.3: AUC-ROC curve to predict 2014 co-consideration network with 6 attributes and 29 attributes

Table 6.1: Confusion matrix in predicting 2013 with 29 features. Average F1-score for 2013 is 0.74. AUC for 2013 train is 0.84 and test is 0.84. True Negative Rate (TNR) and True Positive Rate (TPR) are shown in brackets.

| | | 2013 training prediction | | 2013 test prediction on held-out links | |
|---|---|---|---|---|---|
| | | 0 | 1 | 0 | 1 |
| Actual Class | 0 | 5390 (TNR 53.90%) | 4610 (FPR 46.10%) | 609 (TNR 54.82%) | 502 (FPR 45.18%) |
| | 1 | 592 (FNR 5.92%) | 9408 (TPR 94.08%) | 75 (FNR 6.75%) | 1036 (TPR 93.25%) |

Table 6.2: Confusion matrix in predicting 2014 and 2015 with 29 features. F1-score for 2014 is 0.65 and 0.65 for 2015. AUC for 2014 is 0.80 and 0.80 for 2015

| | | 2014 test prediction on unseen network | | 2015 test prediction on unseen network | |
|---|---|---|---|---|---|
| | | 0 | 1 | 0 | 1 |
| Actual Class | 0 | 42633 (TNR 61.73%) | 26435 (FPR 38.27%) | 45735 (TNR 61.28%) | 28893 (FPR 38.72%) |
| | 1 | 1811 (FNR 15.17%) | 10124 (TPR 84.83%) | 2195 (FNR 16.43%) | 11167 (TPR 83.57%) |
| AUC | | 0.80 | | 0.80 | |
| F1 score | | 0.65 | | 0.65 | |

Table 6.3: Confusion matrix in predicting 2014 with six features and 296 cars for using the GNN method and the ERGM method. F1 score is 0.60 for the GNN model and 0.31 for the ERGM model, and the AUC is 0.78 for the GNN model and 0.68 for the ERGM model.

| | | 2014 prediction class GNN | | 2014 prediction class ERGM | |
|---|---|---|---|---|---|
| | | 0 | 1 | 0 | 1 |
| Actual Class | 0 | 20336 (TNR 54.95%) | 16675 (FPR 45.05%) | 14993 (TNR 40.51%) | 22018 (FPR 59.49%) |
| | 1 | 867 (FNR 13.04%) | 5782 (TPR 86.96%) | 1384 (FNR 20.82%) | 5265 (TPR 79.18%) |
| AUC | | **0.78** | | 0.68 | |
| F1 score | | **0.60** | | 0.31 | |

Table 6.4: Comparing train AUC and test AUC in different years, different models and different sets of attributes. AUC in link prediction. The goal is to predict the entire network (all existing and non-existing edges) in a 0/1 classification task

| Number of attributes | Train AUC (2013) | Test AUC (2014) | Test AUC (2015) | Test AUC (ERGM) |
|---|---|---|---|---|
| 29 attributes | 0.84 | 0.80 | 0.80 | NA |
| Six attributes | 0.81 | 0.78 | NA | 0.68 |

A confusion matrix first measures the prediction performance along with the training performance in Table 6.1. The right-hand part of Table 6.1 shows the confusion matrix of 2013 test prediction on held-out links. It includes four different combinations of predicted and actual classes. The 609 in the top-left cell is the true negative (the model predicted negative, and it was true), and the 502 in the top right is the false positive (the model predicted positive, and it was false). The associated percentages indicate that for all pairs of nodes without link existence (actual class = 0), 54.82% are predicted correctly, whereas 45.18% are not. Meanwhile, the 75 in the bottom left is the false negative (model predicted negative and it was false), and 1036 in the bottom right is the true negative (the model predicted negative and it was true), which suggests that for all pairs of nodes with link existence (actual class = 1), 93.25% are predicted correctly while 6.75% are not. We further calculate other evaluation metrics to quantify classification performance. The F1 score, which measures the test accuracy in an unbalanced class, was *0.74* for the predicted missing links (the range of the F1 score was [0, 1]), while the AUC was *0.84* for both training set and held-out test set. The higher the AUC, the better the model is — it tells how capable the model is when distinguishing between classes. We note that over-fitting is avoided because the AUCs for both the training and test sets are comparable. While the results in Table 2 are promising, they are of less practical usage. This is because a car manufacturer may care less about predicting relationships between cars in the year of survey completion and more about future predictions, enabling them to make strategic design decisions.

**Predicting entire network for the following year** Once the trained model is converged, the learned parameters for the GNN model and the classification model can predict the co-consideration network in the future years. As a test dataset, the car co-consideration network in 2014 is predicted. First, the 2014 car model set, which has an intersection with the 2013 car set and has newly emerged cars, acts as the input of the prediction process without any link information. Then, an approximate adjacency matrix based on the similarities of nodes is generated through the adjacency prediction model. Next, the node features and approximate prediction model are fed into the GNN model, followed by the classification model. The link existence of each pair of nodes is forecasted with a certain probability threshold.

The performance of GNNs in predicting future networks is one of the most important results in this chapter, which is highlighted in Table 6.2. We show the confusion matrix for the predicted 2014 co-consideration network in Table 6.2. Furthermore, we scoped out the AUC-ROC curve (in Fig. 6.3). The overall AUC for this curve is 0.80. To test the repeatability of our results, we conducted ten runs and found all runs to give results between 0.80 and 0.81. Later, we also discuss how the results generalize to networks created using different cut-off values and the number of neighbors.

**Predicting entire network for the year after next** So far, we have predicted the 2014 co-consideration network based on the training data in 2013. However, as 2014 succeeded in 2013, the market structure did not change dramatically. Among 389 cars in 2013 and 403 cars in 2014, there are 296 cars in common. Therefore, to further assess the prediction capability for the model, we predict the 2015 co-consideration network using the trained model (2013 training data) with the car attributes and similarity-based adjacency matrix.

The predicted results are recorded and evaluated in Table 6.2 and Fig. 6.3, where the F1 score is

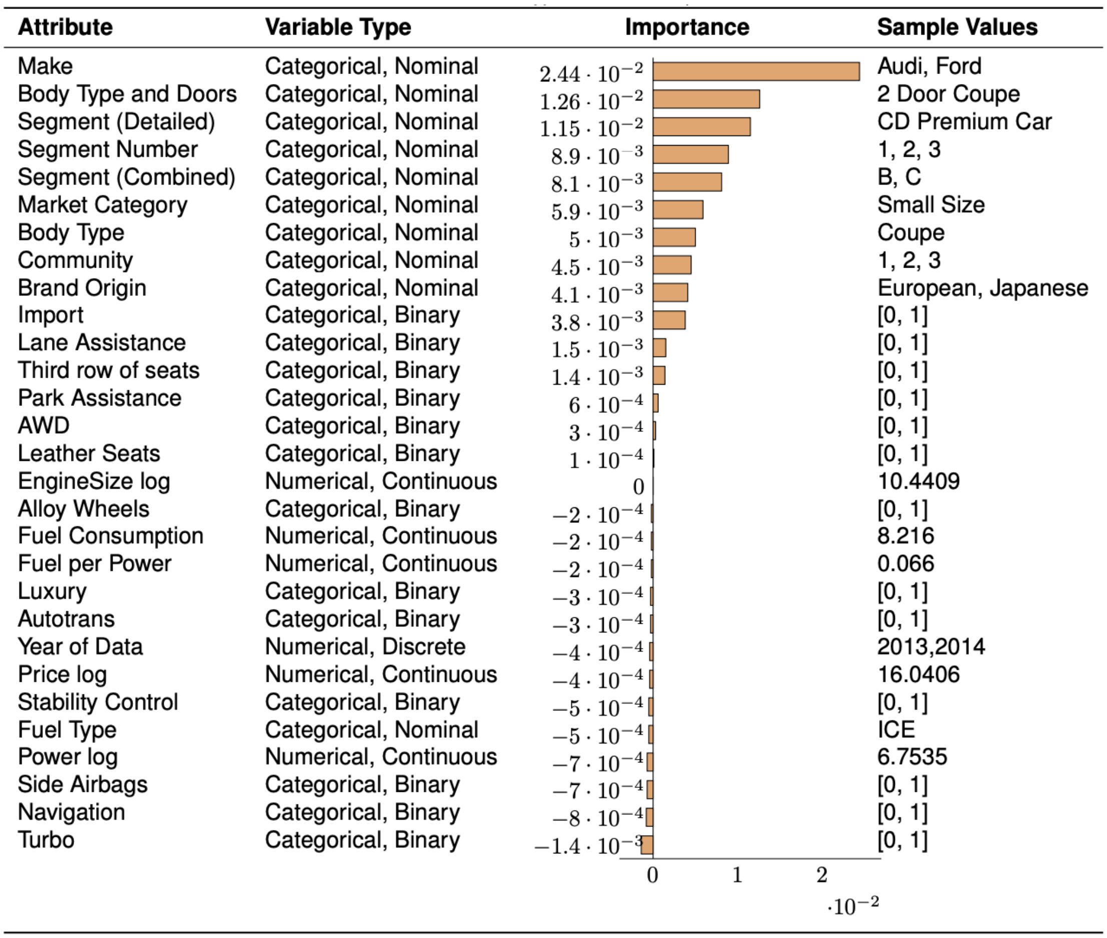
0.65 and AUC is 0.80. Compared to the prediction results in 2014, the prediction in 2015 maintains an equivalent performance, indicating model robustness.

**Comparison with existing statistical network models** In this section, we compare the GNN method with an existing statistical network modeling method — ERGM. In order to make a fair comparison with the literature, we used the same set of input attributes (only six attributes in Chapter 3) and compared the AUC of each model. Besides, as previous studies used a subset of cars and did not predict newly emerged car models, we also took the intersection of 2013 and 2014 cars (296 cars in total) for our analysis.

When only six car features were utilized in the training and prediction model, we found that the prediction results for 2014 data for GNN are significantly better than that of ERGM, as shown in Table 6.3. In the confusion matrix, we observed that in ERGM, the true positive rate (the ratio of true positive to all actual positive) is 79.81% and the true negative rate (the ratio of true negative to all actual negative) is 40.51%. Both of the values are lower than those predicted by GNN. Furthermore, the F1 score of the ERGM is merely 0.31, which is almost half the 0.60 F1 score of the GNN model. The AUC for ERGM prediction is 0.68, which is also less than the corresponding value of 0.78 for the GNN model. All of the evidence suggested that the prediction model of GNN performs better than the traditional statistical network models. It is also important to note that, unlike ERGM, GNN can model a large number of attributes (29 attributes) and unseen data. These benefits prove its effectiveness in modeling networks in comparison to other statistical methods.

### *Interpretability of attributes*

We applied the permutation method to inspect the feature importance to find the decrease in a model score when a single feature value is randomly shuffled. We ran 50 permutations for each



| Attribute | Variable Type | Importance | Sample Values |
|---|---|---|---|
| Make | Categorical, Nominal | $2.44 \cdot 10^{-2}$ | Audi, Ford |
| Body Type and Doors | Categorical, Nominal | $1.26 \cdot 10^{-2}$ | 2 Door Coupe |
| Segment (Detailed) | Categorical, Nominal | $1.15 \cdot 10^{-2}$ | CD Premium Car |
| Segment Number | Categorical, Nominal | $8.9 \cdot 10^{-3}$ | 1, 2, 3 |
| Segment (Combined) | Categorical, Nominal | $8.1 \cdot 10^{-3}$ | B, C |
| Market Category | Categorical, Nominal | $5.9 \cdot 10^{-3}$ | Small Size |
| Body Type | Categorical, Nominal | $5 \cdot 10^{-3}$ | Coupe |
| Community | Categorical, Nominal | $4.5 \cdot 10^{-3}$ | 1, 2, 3 |
| Brand Origin | Categorical, Nominal | $4.1 \cdot 10^{-3}$ | European, Japanese |
| Import | Categorical, Binary | $3.8 \cdot 10^{-3}$ | [0, 1] |
| Lane Assistance | Categorical, Binary | $1.5 \cdot 10^{-3}$ | [0, 1] |
| Third row of seats | Categorical, Binary | $1.4 \cdot 10^{-3}$ | [0, 1] |
| Park Assistance | Categorical, Binary | $6 \cdot 10^{-4}$ | [0, 1] |
| AWD | Categorical, Binary | $3 \cdot 10^{-4}$ | [0, 1] |
| Leather Seats | Categorical, Binary | $1 \cdot 10^{-4}$ | [0, 1] |
| EngineSize log | Numerical, Continuous | $0$ | 10.4409 |
| Alloy Wheels | Categorical, Binary | $-2 \cdot 10^{-4}$ | [0, 1] |
| Fuel Consumption | Numerical, Continuous | $-2 \cdot 10^{-4}$ | 8.216 |
| Fuel per Power | Numerical, Continuous | $-2 \cdot 10^{-4}$ | 0.066 |
| Luxury | Categorical, Binary | $-3 \cdot 10^{-4}$ | [0, 1] |
| Autotrans | Categorical, Binary | $-3 \cdot 10^{-4}$ | [0, 1] |
| Year of Data | Numerical, Discrete | $-4 \cdot 10^{-4}$ | 2013,2014 |
| Price log | Numerical, Continuous | $-4 \cdot 10^{-4}$ | 16.0406 |
| Stability Control | Categorical, Binary | $-5 \cdot 10^{-4}$ | [0, 1] |
| Fuel Type | Categorical, Nominal | $-5 \cdot 10^{-4}$ | ICE |
| Power log | Numerical, Continuous | $-7 \cdot 10^{-4}$ | 6.7535 |
| Side Airbags | Categorical, Binary | $-7 \cdot 10^{-4}$ | [0, 1] |
| Navigation | Categorical, Binary | $-8 \cdot 10^{-4}$ | [0, 1] |
| Turbo | Categorical, Binary | $-1.4 \cdot 10^{-3}$ | [0, 1] |

Figure 6.4: Car attributes type and feature importance

feature in the training data and calculated the drop in performance. These repeats in the process with multiple shuffles were done to ensure accuracy. The results are shown in Figure 6.4. We found that the make of the car, the body type, and the segment are the most critical attributes for the GNN to predict ties.

Figure 6.4 shows that 14 of the 29 attributes have no positive effect on the model prediction. Note that negative values are returned when a random permutation of a feature's values results in a better performance metric than before a permutation is applied. This means the model does not rely on features that have negative values when predicting links for the training data. We observe that most continuous values, such as engine size, price, fuel consumption, and power, are not important. This behavior may either reflect a trend in the data captured by the GNN model or may be caused by a methodology limitation of the applicability of permutation-based methods for mixed (continuous and discrete) data. Understanding the cause of this trend is an exciting direction of research, which can be explored in future work on interpretability analysis.

## 6.3 Link prediction in customer two-stage decision-making process

Building upon the success of GNN methods in the exploration of unidimensional product competition networks, this section seeks to expand the scope of this application to bipartite networks involving customers. The goal is to transition from a simplified network structure to one that more accurately represents the complexities of the real-world market scenario.

In the unidimensional product competition network, we focused primarily on aggregated customer preferences, which were translated into insights regarding product competition. This approach, while beneficial for understanding the broad dynamics of market competition, falls short in capturing the rich heterogeneity and individuality of customer preferences. Recognizing this limitation, we are prompted to introduce customer nodes into the model, thereby advancing our understanding of the interaction between customers and products in the market.

The process of modeling a bipartite network, which essentially involves diversification into two distinct node types - customers and products, offers considerable advantages. It not only bolsters the granularity of our analysis but also poses intriguing challenges to our pre-established methodologies. Accommodating these disparate node types, our model now needs to encapsulate the nuances of the two-stage decision-making process followed by customers – the ’consideration link’ and the ’purchase link’. Recognizing these as distinct yet interconnected facets of consumer behavior, our model seeks to provide a comprehensive portrayal of the customer decision-making journey.

To address this increased complexity, an expansion upon the prior GraphSAGE methodologies becomes necessary. The improved methodology would allow us to accommodate different node types and varying edge types within the same network, thereby enabling us to make more precise predictions and generate more insightful strategic advice. Despite the inherent challenges, it is through the exploration of these complexities that we gain a deeper understanding of customer behavior within the product competition network.

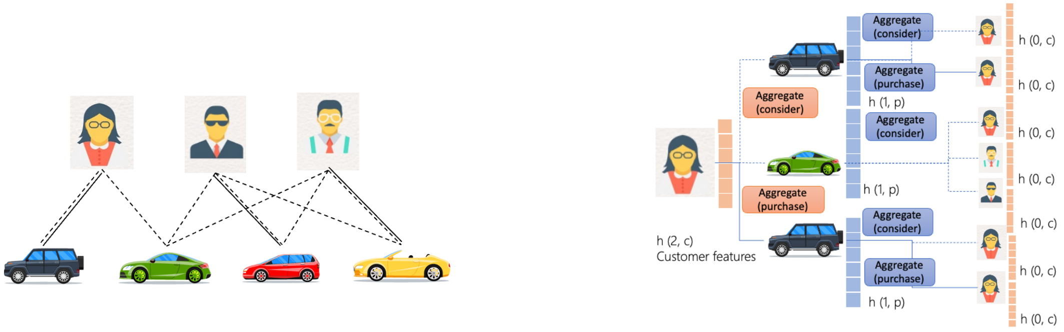
### 6.3.1 Methodology

Link prediction on homogeneous graphs (with one type of nodes and links) using GNN methods can be extended to heterogeneous graphs (with more than one types of nodes and links). HinSAGE (short for Heterogeneous GraphSAGE), an extension of the GraphSAGE method, is an exemplary algorithm for link prediction in heterogeneous graphs (Shang et al., 2016). The primary strength of HinSAGE lies in its ability to accommodate and learn from diverse node and edge types, effectively capturing the intricacies of interactions in customer-product networks.

*Neighbourhood aggregation*

In contrast to homogeneous graphs, where the node representation (for instance, a car node as illustrated in Figure 6.2) is learned and aggregated from its neighboring nodes using a single, fixed, trainable parameter matrix $W_{neighbor}$ in each layer of embedding, HinSAGE produces different parameter matrices depending on the types of nodes and edges. For every unique ordered tuple of $(Node_{type_i}, Edge_{type_k}, Node_{type_j})$, there is a corresponding trainable parameter matrix $W_{ikj}$.

Our application of HinSAGE in the bipartite network involves two discrete node types: customers and products, along with two distinct edge types: considerations and choices. It then separately samples and aggregates neighbor features for these different types of nodes and links. For instance, when learning the embedding from customer nodes (as illustrated in Figure 6.5), in the first layer (1-step neighbor), the weight matrices aggregating the product neighborhood information would involve $W_{neighbor}$ for the consideration link $(Node_{customer}, Edge_{consider}, Node_{car})$ and the purchase link $(Node_{customer}, Edge_{purchase}, Node_{car})$. In the second layer of aggregation (second hop), the product features would be aggregated from the consideration link $W_{neighbor}$ for $(Node_{car}, Edge_{beconsidered}, Node_{customer})$ and the purchase link $(Node_{car}, Edge_{bepurchased}, Node_{customer})$. This process is extended to multiple layers to encapsulate extensive neighborhood information, thereby capturing complex patterns and dependencies in the network.



(a) Bipartite Customer-Product Network

(b) 2-Layer Aggregation of a Center Customer Node

Figure 6.5: Illustration of Neighbourhood Aggregation in HinSAGE

*Edge embedding*

Upon mastering the node embeddings for customers and products, the HinSAGE model computes edge embeddings predicated on these node embeddings. For instance, when there is a ”consideration” link between a customer node and a product node, an edge embedding is derived by applying a function to the embeddings of these nodes. Common functions used for this purpose include the ”inner product” or ”concatenation” operations. This process allows the HinSAGE model to encode both the individual characteristics of each node and the specific relationship between them. By doing so, it captures the essence of the interaction between a customer and a product, leading to a deeper understanding of their interplay within the market ecosystem. It's important to note that the edge embedding in a heterogeneous graph varies depending on the type of edge.

Similarly to the approach employed in link prediction as outlined in section 6.2, the edge embeddings learned through the HinSAGE model can be leveraged for classification tasks. This enables us to predict the presence of a link, thereby enhancing our ability to anticipate customer-product interactions. Therefore, by implementing HinSAGE, we gain the necessary tools to predict customer behaviors.

### 6.3.2 Case study

In this section, our primary focus revolves around a specific task concerning link prediction in heterogeneous graphs. We aim to predict customers' choices based on their consideration sets within a bipartite customer-product network. Previous studies on two-stage modeling have already highlighted the significance of consideration set information in effectively predicting customer choices. Therefore, this research seeks to evaluate the capabilities of Graph Neural Networks (GNNs) in capturing both customers' consideration sets and their subsequent choices.

For each customer node, the GNN incorporates information from both the products they have considered and the products they have purchased. Moreover, the GNN leverages neighborhood information of their connected products, the information is that have been considered or purchased by customers. Consequently, the learned embeddings of customers should include their preferred products and provide valuable market insights.

To conduct our experiments, we utilize car survey data instead of vacuum cleaner survey data, as it offers a more extensive range of information about customers. It enables us to showcase the scalability of GNN models in a more comprehensive manner.

To evaluate the model's performance, we randomly select 10% customers as the testing set. For these testing customers, we remove their purchasing links from the dataset, setting them as empty links. Next, we train the HinSAGE model using the remaining network includes both all customers' training data and training customers' consideration data. The trained model is then utilized to calculate the purchase edge embeddings between the testing customers and the cars they have considered. These edge embeddings capture the relationships between the testing customers and the considered cars. To determine the existence probability of the links, the edge embeddings pass through the classification layer of the model. By calculating the link existence probability, the trained model enables us to predict whether a link, representing the purchase of a car by a testing customer, exists or not.

### *Model settings for HinSAGE*

To implement the HinSAGE model, we begin by defining the initial node embeddings for customers and products. The model adopts a two-layer neural network architecture, with each layer comprising aggregating and updating functions. During the learning process of customer embeddings, the aggregating function in the first layer gathers information from the neighborhood nodes, including the considered and purchased products of a customer node. The updating function then combines this neighborhood information with the initial customer node embedding to generate an updated embedding. This iterative process continues in the second layer, where the aggregating function operates on the updated embeddings from the first layer, capturing more refined information about the node's surroundings. Through this iterative application of aggregating and updating functions, the HinSAGE model learns to encode both the unique characteristics of each node and the relationships between them.

It is worth noting that the HinSAGE model employs a neighborhood sampling strategy to handle computational efficiency. Specifically, in our case, we sample 8 neighbors in the first layer and 4 neighbors in the second layer. If there are fewer neighbors available than the desired number of samples, the algorithm oversamples by repeating the available samples.

Once the node embeddings for customers and products have been obtained, edge embeddings are generated by concatenating the embeddings of connected nodes. This concatenation operation combines the embeddings of a customer node and a product node, resulting in an edge embedding that encapsulates their relationship. These edge embeddings are then connected to a fully connected layer, which functions as a classifier. Leveraging the capabilities of neural networks, the fully connected layer predicts the presence or absence of a link between the nodes.

### *Link prediction results*

In this section, we delve into a thorough examination of the predictive results offered by the GNN model. Our analysis is primarily centered around two core aspects. Firstly, we establish the fact that the GNN model outperforms the ERGM when it comes to prediction accuracy. This superior performance of the GNN model is mainly attributable to its inherent flexibility, a quality that provides it with a distinct edge over ERGM. Our discussion then segues into the second area of focus - the impressive scalability of the GNN model. Unlike ERGM, which struggles to handle larger networks, GNN is highly effective when dealing with larger networks and their accompanying feature sets. In particular, incorporating a broader set of features can significantly enhance prediction accuracy. As a means to illustrate this, we conduct a comparative analysis of the GNN model's performance across two different sizes of input features. we scrutinize the GNN model's capacity to integrate both customer/product features and structural effects, setting it against a widely accepted model used in binary classification prediction. This comparison demonstrates that the GNN model, being network-based, is more effective when incorporating network structural effects. This finding is crucial, reinforcing our key thesis claim concerning the superiority of network-based methods.

**Prediction accuracy compared to ERGM** To scrutinize the predictive capabilities of our models, we contrast the GNN model's performance with that of Exponential Random Graph Models (ERGM). The objective here is to ascertain whether the inherent flexibility of deep neural networks could potentially facilitate superior predictive accuracy. To ensure a level playing field in terms of comparison, we resort to the same dataset as employed in Sha et al. (2023). This dataset comprises 5,000 customers, each having six cars in their consideration set, and includes features such as price, fuel consumption, power, and brand origin.

To measure performance and draw comparisons with the findings from Sha et al. (2023), we employ the Top-N choice probability metric. The essence of Top-N choice probability lies in utilizing the predicted choice probabilities of the top N alternatives (where N can be 1,2,..., h) (h is the number of considerations) within a choice set and juxtaposing these predictions against a customer's final choice. An accurate prediction instance is registered when the predicted choice is encompassed within the Top-N alternatives (Cremonesi et al., 2010).

To thoroughly evaluate the models, we compare not only the Top-N accuracy of the GNN model with the ERGM model but also include variations of the ERGM model both with and without network structural effects. This allows us to establish a robust benchmark for comparison (as depicted in Figure 6.6). Evidently, the GNN model shows a significant performance enhancement when compared to ERGM models, regardless of whether network structural effects are incorporated or not. Notably, the GNN model has the inherent ability to capture network structural effects implicitly. Furthermore, it offers greater flexibility by accounting for more implicit network structures, in contrast to the ERGM model without network structural effects, which only encompasses star effects and degree effects within the model.

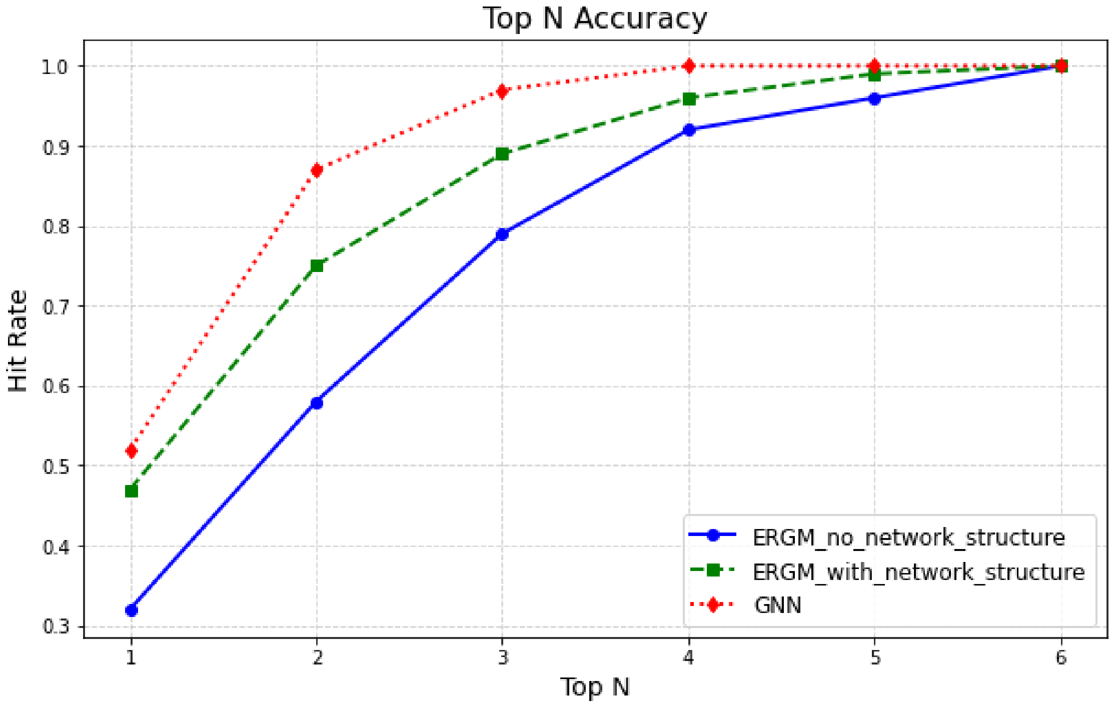
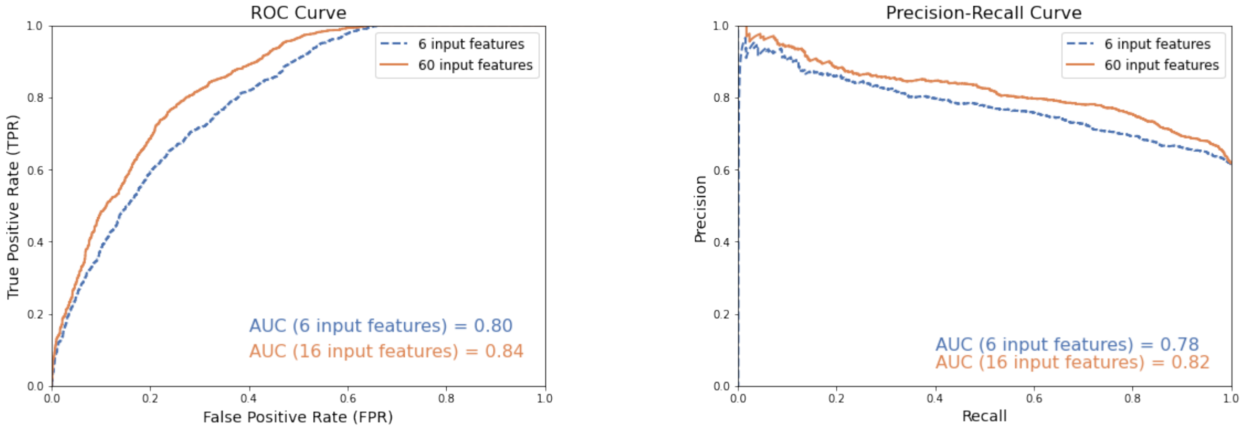


Figure 6.6: Top N accuracy of GNN model, ERGM model and MLP model

**Scalability and feature capture with GNN** Subsequently, we investigate the scalability of GNN methods, with a specific focus on their capacity to accommodate larger networks that encapsulate an increased volume of customers and a more extensive set of features.

Our more extensive dataset is comprised of 18,054 customers from a 2013 car survey. Each customer considered between 1 to 3 cars before finalizing their decision. The data includes 10 customer features (e.g., demographic information, usage-context attributes, and personal viewpoints) and 6 car features (e.g., engineering characteristics like engine size and fuel consumption, as well as customer rating data).

We first evaluate the model performance by incorporating a total of 16 features and then again with a smaller set of features. To assess the performance of our models, we employ several metrics, including the Receiver Operating Characteristic (ROC) curve and the Precision-Recall (PR) curve, as illustrated in Figure 6.7. The ROC curve provides an evaluation of the balance between the true positive rate and the false positive rate, giving a complete overview of the model's performance across different classification thresholds. The PR curve highlights the trade-off between precision and recall, particularly focusing on the prediction of the positive class. These curves are crucial for understanding the distinguishing power of the models and their ability to handle imbalanced data. Importantly, the GNN model with the larger feature set surpasses the others in both the ROC and PR curves, indicating a more accurate reflection of overall customer preferences. In addition to these curves, we also assess prediction accuracy, which measures the model's ability to predict customer preferences based on their consideration sets. The GNN model incorporating more features achieves an accuracy of 77.35% in predicting customer choices among those who considered more than one car. In contrast, the GNN model with fewer features results in a slightly lower accuracy of 70.68%.



(a) AUC values for models with 6 features and 16 features

(b) PR values for models with 6 features and 16 features

Figure 6.7: Comparison of model performance with different number of features

To further verify the efficiency of our model, we compare the model's results with a common classification model, the Multilayer Perceptron (MLP). We feed the MLP models with customer attributes and product attributes, aiming to predict whether a customer will make a purchase within their consideration set. As shown in Figure 6.8, the GNN model incorporating more features again outperforms the MLP in both the ROC and PR curves, indicating a better understanding of overall customer preferences. Additionally, the GNN model achieves a higher prediction accuracy than the MLP model. The GNN model with more features yields an accuracy of 77.35% in predicting customer choices among those who have considered more than one car. In comparison, the GNN model with small feature yields a slightly lower accuracy of 74.32%.

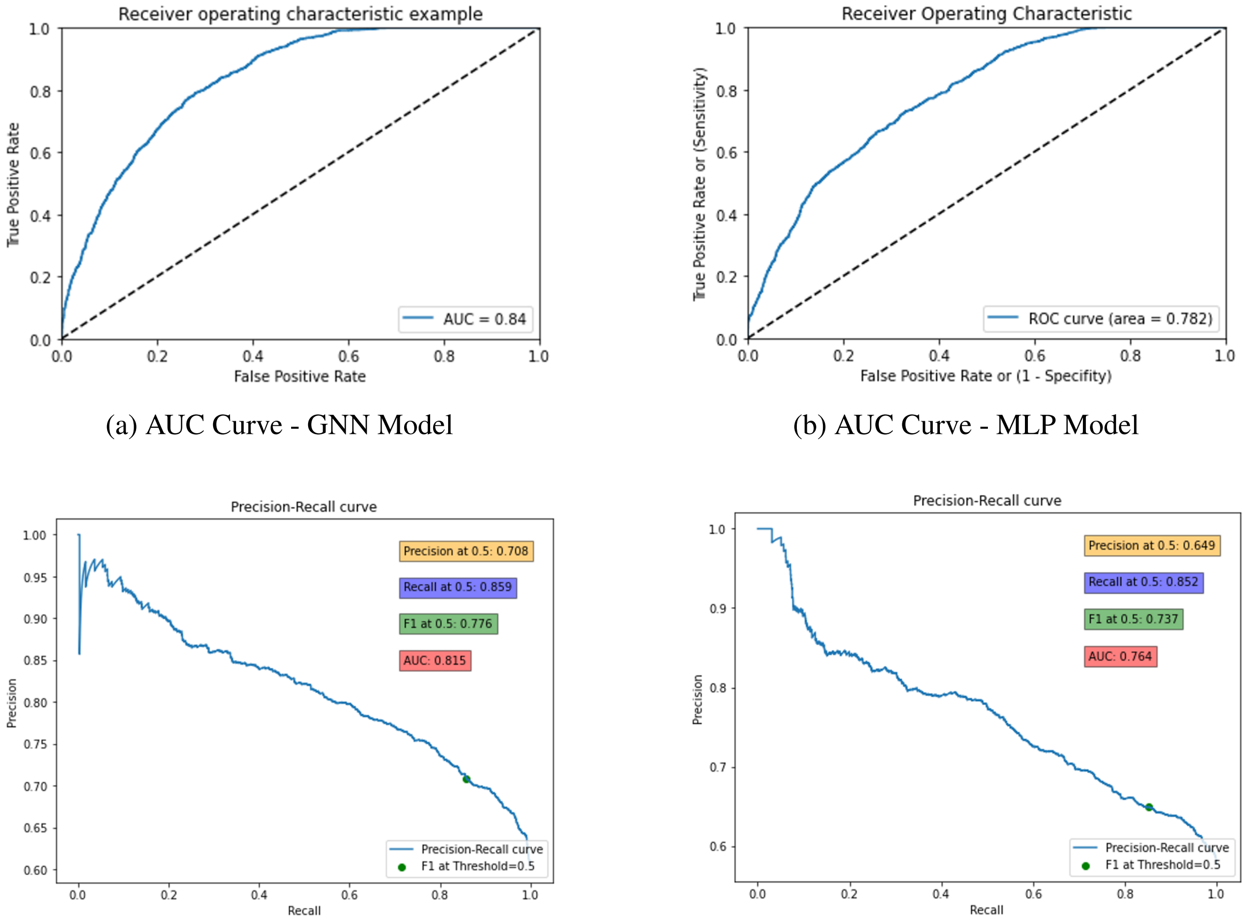
## 6.4 Discussion

In this section, we delve into the advantages and constraints associated with GNN-based approaches in customer-product networks.

### 6.4.1 Product competition network

Car buying decisions hinge on a multitude of factors such as budget, driving needs, and desired features. Manufacturers must thoroughly understand these dynamics to enhance their market share. The proposed Graph Neural Network (GNN) model aids in this process by predicting market competition and identifying potential competitors when introducing or modifying car models. This can help designers strategically plan design changes and new model releases.

The GNN model's effectiveness is evaluated using F1 and AUC scores, indicating a strong prediction capability with an F1 score of 0.60 and over 80% true positive rate. Moreover, it provides insights into feature importance in the co-consideration network, with factors like make, body type, and import playing significant roles. However, these insights require practical validation,



(a) AUC Curve - GNN Model

(b) AUC Curve - MLP Model

(c) PR Curve - GNN Model

(d) PR Curve - MLP Model

Figure 6.8: Comparison of AUC and PR Curves between GNN and MLP Models

considering potential differences between reported customer behaviors and their actual actions.

Nonetheless, the model has limitations. While capable of discerning general market competition patterns, it may struggle when significant shifts in customer preferences occur, as witnessed during major events like the 2020 global pandemic. Also, the model's performance heavily depends on the parameters set by the modeler, which can result in oversensitivity or poor performance under specific settings. Hence, to leverage the model optimally, the underlying network should remain relatively stable, and appropriate parameter settings need to be determined.

### 6.4.2 Customer two-stage decision-making

In our heterogeneous link prediction study, we scrutinized the effectiveness of graph embedding methods for predicting customer purchase behaviors within their consideration set. The application of these methods provides a comprehensive and nuanced understanding of consumer behavior, thereby aiding in the prediction of purchase decisions.

Within the context of the Graph Neural Network (GNN) model, we incorporated both customer and product nodes, as well as consideration and choice links into the network. This strategy enabled us to predict the purchase link. The GNN model proved adept at learning node embeddings based on node attributes and neighbouring nodes. When compared with the Multilayer Perceptron (MLP) model, the GNN model demonstrated superior performance due to its capacity to capture the network structural effects. This ability mirrors the process in which customer behavior is shaped and influenced by the overall market. Further, when the GNN model was compared with the Exponential Random Graph Model (ERGM), it revealed a significantly higher prediction accuracy. This superior performance can largely be attributed to the GNN model's flexible form and its ability to capture implicit network structures.

Despite its strengths, the GNN model exhibits some limitations. First, it lacks interpretability. This limitation restricts our ability to investigate the impact of certain attributes on customers' decision-making processes. Second, network structures in a bipartite network are confined to relationships between customers and products. Consequently, the GNN model may not fully capture some of the interesting effects in the entire market, such as peer influence among customers. This limitation indicates the need for the integration of additional data or the use of other modeling techniques to ensure a comprehensive understanding of market dynamics.

## 6.5 Conclusion

In this chapter, we've elucidated the applications of GNN-based methods in network-based customer preference modeling, exploring both the unidimensional car competition network and the bipartite customer two-stage decision-making network. Our findings affirm the potency of GNN-based methods in effectively encapsulating node representations based on their features and neighbourhood, implicitly reflecting network structural effects.

Notably, the GNN-based methods outperformed traditional network-based statistical models like ERGM, yielding superior prediction accuracy. This illustrates the significant potential of such methods in understanding and predicting customer behavior, offering a robust tool for marketers and strategists.

In the realm of the car competition network, we delved deeper into the feature importance using interpretable machine learning tools. This facilitated an insightful understanding of how various car features play their part in market competition, enriching the overall interpretability of our model.

In essence, these contributions underscore the strength of GNN-based methods in providing a comprehensive and nuanced understanding of customer preferences within a network-based modeling framework. Despite the noted limitations, these methods hold great promise for enhancing predictive accuracy and providing richer insights in the dynamic and complex domain of customer behavior analysis.

# CHAPTER 7
# CONCLUSION AND FUTURE WORK

## 7.1 Contribution of the dissertation

The primary contribution of this work is the development of network-based approaches for understanding customer decision-making processes and product competition within markets. This dissertation builds upon pioneering research on the MCPN (multilevel customer product network) framework, where the distinct roles of ’customers’ and ’products’ are modeled, and multiple types of relations, such as customers’ considerations and choices in the two-stage decision-making process and product competitions, are captured in the network modeling. More specifically, this dissertation focuses on the development of new methods, which include taking the link strength into modeling, integrating market segmentation to capture heterogeneous customer preferences, and introducing graph neural network-based methods for modeling the networks. Moreover, to generalize the application of network-based methods, a systematic information retrieval and survey design approach is proposed for effective data collection, covering the customer decision-making process, customer attributes, their social network relations, and product attributes.

Firstly, **the weighted network approach to product competition analysis** offers a significant advantage in our understanding of the competitive landscape compared to traditional binary network approaches. By assigning link weights within the product co-consideration and choice network, we can capture the competition strength and represent aggregated customer preferences more effectively. Incorporating these weighted links enables us to delve deeper into the intricate relationships and interactions among competing products in the marketplace, providing valuable insights for businesses. This innovative methodology allows us to surpass the limitations of a simple binary representation of product competition, offering a more nuanced and accurate depiction of the competitive dynamics at play. The weighted connections serve a dual purpose: they not only illustrate the presence of competition between products but also convey the intensity of the rivalry. This comprehensive perspective provides a clearer understanding of the competitive environment, highlighting the strengths and weaknesses of different products. Additionally, the superiority of the weighted network approach becomes evident as it provides a more comprehensive explanation of the significance of different product features compared to previous binary network approaches. By considering the level of competition, businesses can pinpoint the key product attributes that influence consumer preferences and, consequently, shape the intensity of rivalry among competing products. This knowledge empowers businesses to make informed decisions concerning product development, marketing strategies, and pricing, ultimately positioning themselves more competitively in the market.

Secondly, in product markets characterized by diverse customer preferences, we have introduced **network-based market segmentation techniques** and integrated them into the bipartite network. This approach involves segmenting the market into distinct groups based on customer heterogeneity and product associations. The proposed methodology has been validated using vacuum cleaner survey data and has proven effective in deciphering how various customers weigh different product attributes. Furthermore, the market segmentation methods outperform models that assume homogeneity in customers' preferences.

Thirdly, with the goal of enhancing link prediction accuracy in both the product competition network (unidimensional product networks) and customers' two-stage decision-making process (bipartite customer-product networks), we have explored **graph neural network approaches**. These approaches learn the node representation by taking into account node attributes and neighboring nodes. Leveraging the computational capability of deep neural networks, the graph-based network modeling approach has demonstrated superior predictive capabilities, with more node attributes and latent network structures being captured by the algorithm. Furthermore, although the graph neural network is a black-box model, the use of an interpretable machine learning tool allows us to discern and rank the importance of various features. Preliminary exploration of the heterogeneous graph has enabled the application of graph neural network approaches to bipartite networks, enabling the prediction of customer choices within their consideration sets. Further exploration in this area is particularly suitable for markets with large volumes of data and the need to learn about embedded customer preferences, a task that outpaces traditional network-molecular approaches, which can only consider a limited number of attributes and network structures.

Last but not least, a significant contribution of this work, and a cornerstone of the network-based customer preference modeling approach, is the development of **a systematic framework for information retrieval and survey design**. This framework is specifically geared towards collecting customer-revealed preference data, along with product attributes, customer attributes, and information about customers' social networks. Our initial survey was designed and launched to collect data on the household vacuum cleaner market. Both the survey design methods and the data collected are open source and available to other researchers and product designers interested in investigating customer preferences in different product markets. Case studies based on the vacuum cleaner data we collected have already been utilized to explore product competition relationships and heterogeneous customer preference modeling. Similar methods have been deployed to study the electric vehicle market in the U.S. The aim of this new survey is to investigate customer preferences for electric vehicles, with a specific focus on how social influence can impact customers' adoption of new technologies.

## 7.2 Limitations and Discussions

Despite the contributions made by the network-based methods mentioned above, this study does have some limitations, and misuse could potentially result in adverse impacts.

**Indirect information in social networks** The study of social networks poses a significant challenge: the lack of direct social influence data. The network-based methods presented are designed to understand the interdependencies among customer decision-making processes. To achieve this, it is essential to construct relationships among customer nodes in the customer layer, as proposed in the Multidimensional Customer Product Network (MCPN). However, the complexity of customer data introduces substantial obstacles to comprehensive network data collection. This process requires detailed information about each individual in the market and their respective social connections. Regrettably, without direct social information, establishing these links within the customer strata remains a considerable challenge.

Efforts are currently being made to tackle this issue. A new survey has been launched to collect data specifically related to individual customers' social networks. Although this does not provide a direct social network among customers, it offers important insights for building models of social influence on technology adoption. This information could potentially enhance our understanding of how social networks impact customers' adoption of new technologies."

**The tradeoff between ERGM and GNN** Our discussion focuses on the trade-offs between two popular methods for analyzing networks: Exponential Random Graph Models (ERGMs) and Graph Neural Networks (GNNs). The ERGM model, a stochastic model, simulates an entire network and determines the optimal parameters via an MCMC process. It provides a platform for testing hypotheses using statistical models, similar to **hypothetico-deductive** methods. However, the ERGM model may be constrained by limitations of nodal attributes, edge attributes, and specific network structural attributes. These restrictions inhibit our ability to investigate product and customer attributes and network structures of interest, imposing limitations on prediction accuracy due to its fixed mathematical form and constrained feature space.

In contrast, Graph Neural Network methods have proven effective at capturing more comprehensive information, considering product attributes, customer attributes, and latent network neighboring effects. However, interpreting the model remains a challenge due to the complexity and non-linearity of the model. Despite these constraints, its ability to simulate or predict customer behaviors provides a platform for **inductive reasoning**, which facilitates theory development based on observations.

The debate surrounding the disparate cultures in employing statistical models to derive conclusions from data originates from the work of Breiman (2001b). This body of work highlights the contrast between data models, which assume that data is generated by given stochastic models, and algorithmic models, which treat the data mechanism as unknown. It draws attention to the potential shortcomings of data models in dealing with complex datasets, and the risk of leading to dubious conclusions. This calls for a deeper exploration of the effective use of both ERGMs (representative of data models) and GNNs (representative of algorithmic models) in various scenarios.

Several potential research directions stem from the trade-offs between ERGM and GNN. Firstly, **abductive reasoning**, which entails generating the most plausible explanation from data and observations (Haig & Haig, 2018; Ren et al., 2018), could potentially bridge the gap between ERGMs and GNNs. Consider using abductive reasoning to enhance the interpretability of GNNs. We can use data and observations to derive plausible explanations for phenomena based on GNNs. If cars with park assistance features are predicted to be more popular among customers, we can hypothesize that park assistance is a significant factor in the decision-making process. Subsequently, the

ERGM model could be used to test the hypothesis for the specific feature's statistical significance.

Secondly, as shown in this dissertation, **interpretable machine learning techniques** make it possible to determine feature importance in a model. These techniques illuminate the roles various attributes play in influencing the predictions made by a model. Notable among interpretable machine learning techniques are SHapley Additive exPlanations (SHAP; (Lundberg & Lee, 2017)), Permutation Feature Importance (PFI; (Breiman, 2001a)), and Partial Dependence Plots (PDPs; (Friedman, 2001)). SHAP, based on game theory, calculates importance values for each feature for a particular prediction, accounting for both individual and interaction effects. PFI assesses feature importance by shuffling one feature at a time and measuring the resultant decrease in model performance, under the notion that significant features induce a noticeable performance drop when shuffled. PDPs are graphical visualizations that illustrate the effect of one or two features on the model's predicted outcome, aiding in understanding the relationship between the target response and selected features. However, it's crucial to understand that while these techniques are powerful, they provide insights different from those obtained through parameter estimation in ERGMs, where the parameters can be used to interpret the significance of product attributes and network structural effects in network link formation.

In summary, we've initiated a discourse on the trade-offs of the network-based modeling approach between interpretability with a fixed and simpler mathematical expression (typical in ERGM) and prediction accuracy with a more flexible but black-box model (common in GNN). Both researchers and product designers must carefully select the most effective models or use a combination of different models based on their need for product feature interpretation or predicting future market evolution.

**Balancing Act: Ethical and Practical Challenges of Data-Driven Network-Based Customer Preference Modeling in Design Research** In light of the increasing integration of data-driven methodologies into the design landscape, critical examination of the implications of algorithm-based methods is imperative (Kellogg et al., 2020). The widespread use of these methods in design research raises several important concerns that require careful examination.

Firstly, while data-driven design often focuses on **profit and economic value**, it's crucial not to lose sight of human factor. Buchanan (2001) asserts that the design holds the power to solve human problems and address societal needs. However, in the context of network-based models geared towards maximizing profits via market share expansion and cost reduction, there's a risk of not paying enough attention to aspects like human welfare, emotions, and societal benefits in the product design process. Essentially, design should also be about people, not just profits.

Secondly, while advancements in network-based models have shown promise in yielding more accurate results, the inherent **limitations of algorithmic intelligence** remain a concern. Designers may struggle when data-driven models contradict their intuitive understanding. For instance, in Chapter 5's network-based model for understanding customer preferences towards vacuum cleaners, without applying market segmentation, the model suggested customers prefer vacuum cleaners with lower suction power. The finding changes when we segment customers into price-sensitive and innovation-passionate groups. We see that only price-sensitive customers prioritize low cost and do not emphasize suction power in their decision-making process. Mis-specified models like these can hinder the provision of effective design insight. Therefore, building **trust** in algorithmic tools needs to be done cautiously, ensuring a balance between data insights and expert intuition.

Thirdly, data collection and analysis methods could unintentionally introduce **biases** with significant implications. Noble (2018) discussed how biases in data can lead to discrimination, especially regarding gender and race. The issue is magnified by a lack of diversity and representativeness in data collection, which necessitates rigorous data management to reduce these biases. When launching products to a specific target market, examining demographic and social attributes might inadvertently introduce biases and discrimination into the design process.

Fourthly, the relentless pursuit of customer data, which is integral to algorithm-based design, raises the issue of **data privacy**. The performance of data-driven network-based models largely hinges on exhaustive, high-volume data, but this intensive data-tracking could lead to customers feeling their privacy has been invaded. This exacerbates the ethical quandaries surrounding data-driven methodologies and necessitates the adoption of stringent data privacy norms.

Lastly, while algorithms are meant to support designers, they might end up taking over some of the design jobs because of automation. Autor (2015) explored the impact of automation on employment and stressed the need for examining its effects carefully. The design research community needs to think about how algorithms and humans can work together in harmony.

In conclusion, as data-driven, algorithm-based methods become more popular in design, it's important to think about their wider impact. Balancing **economic goals** with human values, being careful about **algorithm limitations**, tackling **data biases**, respecting **customer privacy**, and considering **the future of design jobs** are all crucial steps in responsibly advancing design methodologies.

## 7.3 Opportunities for future research

The findings of this study offer valuable insights into the use of network-based models for investigating product competition and customer preferences. However, there are several avenues for further research that could build upon these insights.

One promising direction for future study is the investigation of **social influence** on customer preference modeling. Specifically, this involves delving into how an individual's likes, dislikes, or inclinations towards certain product features or even entire products can be shaped or modified by the attitudes or preferences of others in their social circle. For this purpose, researchers can employ a statistical framework known as Autologistic Actor Attribute Models (ALAAM) (Daraganova & Robins, 2012). ALAAM provides a robust mechanism to examine the interaction between social influence and individual preferences by considering the dependencies among individuals within a social network. This refers to the interconnected relationships and interactions people have within their social group, and how these may impact their attitudes and behaviors. By leveraging ALAAM, researchers can analyze the extent to which individuals' choices and preferences are influenced by the attitudes and preferences of their social connections. With the growing availability of survey data currently being collected in the U.S. car market, researchers can examine how social factors shape individual preferences and decision-making processes. The collected data reveal both the attitudes and preferred features related to car selection of the respondents (who we refer to as ”egos”) as well as those of their connections in their social network (whom we refer to as ”alters”). Researchers can utilize such data to examine the ways in which social factors – such as peer influence, societal norms, or group dynamics – contribute to shaping individual preferences and the decision-making process. This could uncover previously hidden patterns or trends, thereby providing more insight into consumer behavior. By analyzing the role and impact of social networks and social influence in guiding consumer behavior, researchers have the potential to gain a deeper understanding of consumer preference formation. This would offer valuable insights into the decision-making process of consumers, which in turn could be leveraged to design more effective, targeted product design and marketing strategies. These strategies could take into account not just the individual's personal preferences, but also the social influences that impact these preferences, thereby leading to more successful and impactful marketing efforts.

Another area primed for deeper exploration revolves around the use of **heterogeneous graph**

**neural networks (HGNNs) within multidimensional customer-product networks.** These models can take into account the complex interrelationships between customers, products, and decision-making processes. An initial investigation into the application of heterogeneous GNNs in a bipartite network has been conducted, as detailed in Chapter 6. This study leveraged the customers' consideration and choice behavior, proving the effectiveness of heterogeneous GNNs in processing networks with diverse node and edge types. However, there is potential for further research to expand upon this foundation, extending the network analysis to incorporate more comprehensive information. A key feature of heterogeneous GNNs is their ability to learn the node embedding mechanism. This process distills nodes – which could represent customers or products – into lower-dimensional, numerical representations known as latent node embeddings. These embeddings retain essential network structure and attribute information, effectively capturing the relationships and characteristics of nodes. Through a detailed analysis of these latent node embeddings, researchers have the potential to uncover previously hidden insights into customer behavior and product competition. This richer understanding could ultimately lead to the development of more accurate and predictive models for deciphering consumer preferences and behavior. Such models could revolutionize businesses' abilities to anticipate customer needs and preferences, thereby driving the development of more effective product strategies and targeted marketing campaigns.

In addition to the previously mentioned research directions, another promising area for future study lies in the **analysis and interpretation of latent node embeddings** in GNN–based models. As the complexity and sophistication of these models increase, it becomes increasingly crucial to understand not only the output but also the internal mechanisms used by the model to represent the underlying data and make predictions. Latent node embeddings are a crucial part of this process. They distill the information of the original nodes (such as customers and products) into a lower-dimensional space. These embeddings, which retain the essential network structure and attributes, can be thought of as the ’knowledge’ the model has learned from the data. By gaining a deeper understanding of how the model is representing the data, researchers can garner insights that aren’t immediately apparent from the raw data or the model’s output. This could include the detection of unexpected patterns, the identification of influential nodes, or a better understanding of the relationship between different nodes. Ultimately, by focusing on the interpretability of the node embeddings, researchers can develop models that are not only more accurate but also more comprehensible. These models would provide a clearer understanding of the phenomena being studied, enhancing the value of the insights derived and making the models more usable for practitioners in the field.

# REFERENCES

Ahmed, F., Cui, Y., Fu, Y., & Chen, W. (2021). A graph neural network approach for product relationship prediction. *ASME 2021 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*.

Akai, R., Amaya, H., & Fujita, K. (2010). *Product Family Deployment Through Optimal Resource Allocation Under Market System* (Vol. Volume 1: 36th Design Automation Conference, Parts A and B).

Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: A survey. *Data Mining and Knowledge Discovery*, *29*(3), 626–688.

Albert, R., Jeong, H., & Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *nature*, *406*(6794), 378.

Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of econometrics*, *89*(1-2), 57–78.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347.

An, W. (2016). Fitting ERGMs on big networks. *Social Science Research*, *59*, 107–119
Special issue on Big Data in the Social Sciences.

Aral, S., & Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management science*, *57*(9), 1623–1639.

Argo, J. J. (2020). A contemporary review of three types of social influence in consumer psychology. *Consumer Psychology Review*, *3*(1), 126–140.

Atwood, J., & Towsley, D. (2015). Diffusion-convolutional neural networks. *arXiv preprint arXiv:1511.02136*.

Autor, D. H. (2015). Why are there still so many jobs? the history and future of workplace automation. *Journal of Economic Perspectives*, *29*(3), 3–30.

Bao, Q., Sinitskaya, E., Gomez, K. J., MacDonald, E. F., & Yang, M. C. (2020). A human-centered design approach to evaluating factors in residential solar pv adoption: A survey of homeowners in california and massachusetts. *Renewable Energy*, *151*, 503–513.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.

Barnard, A. S., Louviere, J. J., Wei, E., & Zadorin, L. (2016). Using hypothetical product configurators to measure consumer preferences for nanoparticle size and concentration in sunscreens. *Design Science*, *2*.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Beane, T., & Ennis, D. (1987). Market Segmentation: A Review [Publisher: MCB UP Ltd]. *European Journal of Marketing*, *21*(5), 20–42.

Beldona, S., Morrison, A. M., & O'Leary, J. (2005). Online shopping motivations and pleasure travel products: A correspondence analysis. *Tourism Management*, *26*(4), 561–570.

Ben-Akiva, M., Morikawa, T., & Shiroishi, F. (1992). Analysis of the reliability of preference ranking data. *Journal of Business Research*, *24*(2), 149–164.

Ben-Akiva, M. E., & Lerman, S. R. (1985). *Discrete choice analysis: Theory and application to travel demand* (Vol. 9). MIT press.

Bernard, H. (2013). *Methods in human geography* (R. Flowerdew & D. Martin, Eds.). Routledge.

Berry, M. J. (2004). A and GS Linoff. *Data Mining Techniques for Marketing, Sales and Customer Relationship Management*.

Bi, Y., Xie, J., Sha, Z., Wang, M., Fu, Y., & Chen, W. (2018). Modeling Spatiotemporal Heterogeneity of Customer Preferences in Engineering Design. *ASME 2018 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*.

Bi, Y., Qiu, Y., Sha, Z., & Wang, W. (2021). Modeling multi-year customers' considerations and choices in china's auto market using two-stage bipartite network analysis. *Networks and Spatial Economics*, *21*, 365–385.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Box, G. E., & Tiao, G. C. (2011). *Bayesian inference in statistical analysis* (Vol. 40). John Wiley & Sons.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59.

Braha, D., Suh, N., Eppinger, S., Caramanis, M., & Frey, D. (2006). Complex engineered systems. In *Unifying Themes in Complex Systems* (pp. 227–274). Springer.

Breiman, L. (2001a). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L. (2001b). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Brock, W. A., & Durlauf, S. N. (2001). Discrete choice with social interactions. *The Review of Economic Studies*, *68*(2), 235–260.

Buchanan, R. (2001). Design Research and the New Learning. *Design Issues*, *17*(4), 3–23.

Burnap, A., Pan, Y., Liu, Y., Ren, Y., Lee, H., Gonzalez, R., & Papalambros, P. Y. (2016). Improving design preference prediction accuracy using feature learning. *Journal of Mechanical Design*, *138*(7), 71404.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding aic and bic in model selection. *Sociological methods & research*, *33*(2), 261–304.

Campbell, K. E., & Lee, B. A. (1991). Name generators in surveys of personal networks. *Social networks*, *13*(3), 203–221.

Cao, S., Lu, W., & Xu, Q. (2016). Deep neural networks for learning graph representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1).

Cao, W., Robinson, T., Hua, Y., Boussuge, F., Colligan, A. R., & Pan, W. (2020). Graph representation of 3d cad models for machining feature recognition with deep learning. *ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*.

Chatterjee, R. A., & Eliashberg, J. (1990). The innovation diffusion process in a heterogeneous population: A micromodeling approach. *Management science*, *36*(9), 1057–1079.

Chen, H. Q., Honda, T., & Yang, M. C. (2012). An approach for revealed consumer preferences for technology products: A case study of residential solar panels. *ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 379–390.

Chen, H. Q., Honda, T., & Yang, M. C. (2013). Approaches for identifying consumer preferences for the design of technology products: A case study of residential solar panels. *Journal of Mechanical Design*, *135*(6).

Chen, W., Ahmed, F., Cui, Y., Sha, Z., & Contractor, N. (2020). Data-driven preference modelling in engineering systems design. In A. Maier, J. Oehmen, & P. E. Vermaas (Eds.), *Handbook of engineering systems design* (pp. 1–34). Springer International Publishing.

Chen, W., Hoyle, C., & Wassenaar, H. J. (2013). *Decision-based design: Integrating consumer preferences into engineering design*. Springer Science; Business Media.

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*(6).

Conte, D., Foggia, P., Sansone, C., & Vento, M. (2004). Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, *18*(03), 265–298.

Cook, H. E., & DeVor, R. E. (1991). On competitive manufacturing enterprises I: The S-model and the theory of quality. *Manufacturing Review*, *4*(2), 96–105.

Cremonesi, P., Koren, Y., & Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. *Proceedings of the Fourth ACM Conference on Recommender Systems*, 39–46.

Daraganova, G., & Robins, G. (2012). Autologistic actor attribute models. In D. Lusher, J. Koskinen, & G. Robins (Eds.), *Exponential random graph models for social networks: Theory, methods, and applications* (pp. 102–114). Cambridge University Press.

de Nooy, W. (2003). Fields and networks: Correspondence analysis and social network analysis in the framework of field theory. *Poetics*, *31*(5), 305–327.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Draganska, M., & Jain, D. C. (2006). Consumer preferences and product-line pricing strategies: An empirical analysis. *Marketing Science*, *25*(2), 164–174.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68–77.

Fan, W., Ma, Y., Li, Q., He, Y., Zhao, E., Tang, J., & Yin, D. (2019). Graph neural networks for social recommendation. *The world wide web conference*, 417–426.

Farinosi, F., Giupponi, C., Reynaud, A., Ceccherini, G., Carmona-Moreno, C., De Roo, A., Gonzalez-Sanchez, D., & Bidoglio, G. (2018). An innovative approach to the assessment of hydro-political risk: A spatially explicit, data driven indicator of hydro-political issues. *Global Environmental Change*, *52*, 286–313.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3-5), 75–174.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Frischknecht, B. D., Whitefoot, K., & Papalambros, P. Y. (2010). On the Suitability of Econometric Demand Models in Design for Market Systems. *Journal of Mechanical Design*, *132*(12), 121007.

Fu, J., Sha, Z., Huang, Y., Wang, M., Fu, Y., & Chen, W. (2017). Two-stage modeling of customer choice preferences in engineering design using bipartite network analysis. *ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*.

Fu, X., Luo, J.-D., & Boos, M. (2017). *Social Network Analysis: Interdisciplinary Approaches and Case Studies*. CRC Press.

Gaskin, S., Evgeniou, T., Bailiff, D., & Hauser, J. (2007). Two-stage models: Identifying non-compensatory heuristics for the consideration set then adaptive polyhedral methods within the consideration set. *Proceedings of the Sawtooth Software Conference*.

Gerth, R. J., Burnap, A., & Papalambros, P. (2012). *Crowdsourcing: A primer and its implications for systems engineering* (tech. rep.). MICHIGAN UNIV ANN ARBOR.

Gorsuch, R. L. (1983). *Factor analysis*. Lawrence Erlbaum Associates, Hillsdale.

Goyat, S. (2011). The basis of market segmentation: A critical review of literature. *European Journal of Business and Management*, *3*(9), 45–54.

Green, P. E. (1970). *Multidimensional scaling and related techniques in marketing analysis*. Allyn; Bacon.

Green, P. E., Carmone, F. J., & Wachspress, D. P. (1976). Consumer segmentation via latent class analysis. *Journal of Consumer Research*, *3*(3), 170–174.

Green, P. E., & Krieger, A. M. (1991). Product design strategies for target-market positioning. *Journal of Product Innovation Management*, *8*(3), 189–202.

Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of marketing*, *54*(4), 3–19.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of consumer research*, *5*(2), 103–123.

Green, P. E., & Wind, Y. (1975). New ways to measure consumer judgments.

Greenacre, M., & Blasius, J. (2006). *Multiple correspondence analysis and related methods*. CRC press.

Grover, A., & Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864.

Haaijer, R., Wedel, M., Vriens, M., & Wansbeek, T. (1998). Utility covariances and context effects in conjoint mnp models. *Marketing Science*, *17*(3), 236–252.

Haig, B. D., & Haig, B. D. (2018). An abductive theory of scientific method. *Method matters in psychology: Essays in applied philosophy of science*, 35–64.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, 1024–1034.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., Bender-deMoll, S., & Morris, M. (2019). Statnet: Software Tools for the Statistical Analysis of Network Data.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., Morris, M., Wang, L., Li, K., Bender-deMoll, S., & Krivitsky, M. P. N. (2015). Package ‘ergm’.

Hauser, J. R., Ding, M., & Gaskin, S. P. (2009). Non-compensatory (and compensatory) models of consideration-set decisions. *2009 Sawtooth Software Conference Proceedings, Sequin WA*.

Hauser, J. R., & Wernerfelt, B. (1990). An evaluation cost model of consideration sets. *Journal of consumer research*, *16*(4), 393–408.

He, L., Chen, W., Hoyle, C., & Yannou, B. (2012). Choice modeling for usage context-based design. *Journal of Mechanical Design*, *134*(3), 31007.

He, L., Wang, M., Chen, W., & Conzelmann, G. (2014). Incorporating social impact on new product adoption in choice modeling: A case study in green vehicles. *Transportation Research Part D: Transport and Environment*, *32*, 421–434.

He, R., & Zheng, T. (2013). Estimation of exponential random graph models for large social networks via graph limits. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 248–255.

Hoffman, D. L., & Franke, G. R. (1986). Correspondence analysis: Graphical representation of categorical data in marketing research. *Journal of Marketing Research*, *23*(3), 213–227.

Holling, C. S. (2001). Understanding the complexity of economic, ecological, and social systems. *Ecosystems*, *4*(5), 390–405.

Hoyle, C., Chen, W., Ankenman, B., & Wang, N. (2009). Optimal Experimental Design of Human Appraisals for Modeling Consumer Preferences in Engineering Design [071008]. *Journal of Mechanical Design*, *131*(7).

Hoyle, C., Chen, W., Wang, N., & Koppelman, F. S. (2010). Integrated Bayesian hierarchical choice modeling to capture heterogeneous consumer preferences in engineering design. *Journal of Mechanical Design*, *132*(12), 121010.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, *24*(3), 1–29.

Jain, A., Liu, I., Sarda, A., & Molino, P. (2019). Food Discovery with Uber Eats: Using Graph Learning to Power Recommendations [[Online; accessed 01-March-2021]].

Jiang, D., Wu, Z., Hsieh, C.-Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., & Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *Journal of cheminformatics*, *13*(1), 1–23.

John, L. K., Kim, T., & Barasz, K. (2018). Targeting ads without creeping out your customers. *Harvard Business Review*, *96*(1), 62–69.

Johnson, R. (2011). *Multiple discriminant analysis: Marketing research applications*. Marketing Classics Press.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (Vol. 5). Prentice hall Upper Saddle River, NJ.

Kamakura, W. A., Kim, B.-D., & Lee, J. (1996). Modeling preference and structural heterogeneity in consumer choice. *Marketing Science*, *15*(2), 152–172.

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*.

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal*, *17*(6), 441–458.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Krivitsky, P. N. (2012). Exponential-family random graph models for valued networks. *Electronic journal of statistics*, *6*, 1100.

Krivitsky, P. N., & Butts, C. T. (2013). Modeling valued networks with statnet. *The Statnet Development Team*, 2013.

Kumar, D., Chen, W., & Simpson, T. W. (2009). A market-driven approach to product family design. *International Journal of Production Research*, *47*(1), 71–104.

Kumar, D., Hoyle, C., Chen, W., Wang, N., Gomez-Levi, G., & Koppelman, F. S. (2009). Incorporating Customer Preferences and Market Trends in Vehicle Package Design, 571–580.

Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research*, *48*(5), 881–894.

Li, K., Gao, Y., Zheng, H., & Tan, J. (2021). A Data-Driven Methodology to Improve Tolerance Allocation Using Product Usage Data [071101]. *Journal of Mechanical Design*, *143*(7).

Liben-Nowell, D., & Kleinberg, J. (2007). Link prediction in social networks. *Social Networks*, 131–157.

Lilien, G. L., Kotler, P., & Moorthy, K. S. (1995). *Marketing models*. Prentice Hall.

Lin, C.-F. (2002). Segmenting customer brand preference: Demographic or psychographic [Publisher: MCB UP Ltd]. *Journal of Product & Brand Management*, *11*(4), 249–268.

Liu, J., Liao, X., Huang, W., & Liao, X. (2019). Market segmentation: A multiple criteria approach combining preference analysis and segmentation decision. *Omega*, *83*, 1–13.

Louviere, J. J., Fox, M. F., & Moore, W. L. (1993). Cross-task validity comparisons of stated preference choice models. *Marketing Letters*, *4*(3), 205–213.

Louviere, J. J., Hensher, D. A., Swait, J. D., & Adamowicz, W. (2000). *Stated choice methods: Analysis and applications*. Cambridge University Press.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 4765–4774.

Lusher, D., Koskinen, J., & Robins, G. (2013). *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.

MacDonald, E. F., Gonzalez, R., & Papalambros, P. Y. (2009). Preference Inconsistency in Multidisciplinary Design Decision Making. *Journal of Mechanical Design*, *131*(3), 031009.

Malak, R. J., & Paredis, C. J. (2010). Using support vector machines to formalize the valid input domain of predictive models in systems design problems. *Journal of Mechanical Design*, *132*(10).

Marino, K., Salakhutdinov, R., & Gupta, A. (2016). The more you know: Using knowledge graphs for image classification. *arXiv preprint arXiv:1612.04844*.

Mark, T. L., & Swait, J. (2004). Using stated preference and revealed preference modeling to evaluate prescribing decisions. *Health economics*, *13*(6), 563–573.

Matin, S., Farahzadi, L., Makaremi, S., Chelgani, S. C., & Sattari, G. (2018). Variable selection and prediction of uniaxial compressive strength and modulus of elasticity by random forest. *Applied Soft Computing*, *70*, 980–987.

Merino-Castello, A. (2003). Eliciting consumers preferences using stated preference discrete choice models: Contingent ranking versus choice experiment. *UPF economics and business working paper*, (705).

Michalek, J. J., Ceryan, O., Papalambros, P. Y., & Koren, Y. (2006). Balancing marketing and manufacturing objectives in product line design. *Journal of Mechanical Design*, *128*(6), 1196–1204.

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning. *arXiv preprint 2010.09337*.

Narayan, V., Rao, V. R., & Saunders, C. (2011). How peer influence affects attribute preferences: A bayesian updating mechanism. *Marketing Science*, *30*(2), 368–384.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied linear statistical models.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, *69*, 026113.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167–256.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

Parraguez, P., Maier, A., et al. (2017). Data-driven engineering design research: Opportunities using open data. *DS 87-7 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 7: Design Theory and Research Methodology, Vancouver, Canada, 21-25.08. 2017*, 041–050.

Peltier, J. W., & Schribrowsky, J. A. (1997). The use of need-based segmentation for developing segment-specific direct marketing strategies. *Journal of Direct Marketing*, *11*(4), 53–62.

Pescher, C., & Spann, M. (2014). Relevance of actors in bridging positions for product-related information diffusion. *Journal of Business Research*, *67*(8), 1630–1637.

Pilny, A., & Atouba, Y. (2018). Modeling valued organizational communication networks using exponential random graph models. *Management Communication Quarterly*, *32*(2), 250–264.

Putin, E., Mamoshina, P., Aliper, A., Korzinkin, M., Moskalev, A., Kolosov, A., Ostrovskiy, A., Cantor, C., Vijg, J., & Zhavoronkov, A. (2016). Deep biomarkers of human aging: Application of deep neural networks to biomarker development. *Aging (Albany NY)*, *8*(5), 1021.

Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., & Tang, J. (2018). Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. *Proceedings of the eleventh ACM international conference on web search and data mining*, 459–467.

Rai, R. (2012). Identifying key product attributes and their importance levels from online customer reviews. *ASME 2012 international design engineering technical conferences and computers and information in engineering conference*, 533–540.

Rana, T. A., & Cheah, Y.-N. (2017). A two-fold rule-based model for aspect extraction. *Expert Syst. Appl.*, *89*(100), 273–285.

Reichardt, J., & Bornholdt, S. (2006). Statistical mechanics of community detection [Publisher: American Physical Society]. *Physical Review E*, *74*(1), 016110.

Ren, Y., & Papalambros, P. Y. (2011). A design preference elicitation query as an optimization process. *Journal of Mechanical Design*, *133*(11).

Ren, Y., Cedeno-Mieles, V., Hu, Z., Deng, X., Adiga, A., Barrett, C., Ekanayake, S., Goode, B. J., Korkmaz, G., Kuhlman, C. J., et al. (2018). Generative modeling of human behavior and social interactions using abductive analysis. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 413–420.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). ‘why should i trust you?’ explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Roberts, J. H., & Lattin, J. M. (1991). Development and testing of a model of consideration set composition. *Journal of Marketing Research*, *28*(4), 429–440.

Sawhney, M., Verona, G., & Prandelli, E. (2005). Collaborating to create: The internet as a platform for customer engagement in product innovation. *Journal of Interactive Marketing*, *19*(4), 4–17.

Scott, T. A. (2016). Analyzing policy networks using valued exponential random graph models: Do government-sponsored collaborative groups enhance organizational networks? *Policy Studies Journal*, *44*(2), 215–244.

Sha, Z., Huang, Y., Fu, S., Wang, M., Fu, Y., Contractor, N., & Chen, W. (2018). A Network-Based Approach to Modeling and Predicting Product Co-Consideration Relations. *Complexity*, *2018*.

Sha, Z., Bi, Y., Wang, M., Stathopoulos, A., Contractor, N., Fu, Y., & Chen, W. (2019). Comparing utility-based and network-based approaches in modeling customer preferences for engineering design. *Proceedings of the Design Society: International Conference on Engineering Design*, *1*, 3831–3840.

Sha, Z., Cui, Y., Xiao, Y., Stathopoulos, A., Contractor, N., Fu, Y., & Chen, W. (2023). A network-based discrete choice model for decision-based design. *Design Science*, *9*, e7.

Sha, Z., Huang, Y., Fu, J. S., Wang, M., Fu, Y., Contractor, N., & Chen, W. (2018). A network-based approach to modeling and predicting product coconsideration relations. *Complexity*, *2018*.

Sha, Z., Moolchandani, K., Panchal, J. H., & DeLaurentis, D. A. (2016). Modeling Airlines' Decisions on City-Pair Route Selection Using Discrete Choice Models. *Journal of Air Transportation*.

Sha, Z., & Panchal, J. H. (2014). Estimating Local Decision-Making Behavior in Complex Evolutionary Systems. *Journal of Mechanical Design*, *136*(6), 61003.

Sha, Z., Saeger, V., Wang, M., Fu, Y., & Chen, W. (2017). Analyzing customer preference to product optional features in supporting product configuration. *SAE International Journal of Materials and Manufacturing*, *10*(3), 320–332.

Sha, Z., Wang, M., Huang, Y., Contractor, N., Fu, Y., & Chen, W. (2017). Modeling product co-consideration relations: A comparative study of two network models. *Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 6: Design Information and Knowledge, Vancouver, Canada, 21-25.08. 2017*.

Shalizi, C. R., & Rinaldo, A. (2013). Consistency under sampling of exponential random graph models. *Annals of statistics*, *41*(2), 508.

Shang, J., Qu, M., Liu, J., Kaplan, L. M., Han, J., & Peng, J. (2016). Meta-path guided embedding for similarity search in large-scale heterogeneous information networks. *arXiv preprint arXiv:1610.09769*.

Shao, W. (2007). *Consumer Decision-Making: An Empirical Exploration of Multi-Phased Decision Processes*. Griffith University.

Shin, J., & Ferguson, S. (2017). Exploring product solution differences due to choice model selection in the presence of noncompensatory decisions with conjunctive screening rules. *Journal of Mechanical Design*, *139*(2).

Shocker, A. D., Ben-Akiva, M., Boccara, B., & Nedungadi, P. (1991). Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing letters*, *2*(3), 181–197.

Silk, M. J., Weber, N. L., Steward, L. C., Hodgson, D. J., Boots, M., Croft, D. P., Delahay, R. J., & McDonald, R. A. (2018). Contact networks structured by sex underpin sex-specific epidemiology of infection. *Ecology letters*, *21*(2), 309–318.

Simon, H. A. (1977). The organization of complex systems. In *Models of discovery* (pp. 245–261). Springer.

Simons, R. (2014). Choosing the right customer. *Harvard Business Review*, *92*(3), 48–55.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological methodology*, *36*(1), 99–153.

Stankevich, A. (2017). Explaining the consumer decision-making process: Critical literature review. *Journal of International Business Research and Marketing*, *2*(6).

Stivala, A., Robins, G., & Lomi, A. (2020). Exponential random graph model parameter estimation for very large directed networks. *PloS one*, *15*(1), e0227804.

Stivala, A. D., Gallagher, H. C., Rolls, D. A., Wang, P., & Robins, G. L. (2020). Using sampled network data with the autologistic actor attribute model. *arXiv preprint arXiv:2002.00849*.

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., et al. (2020). A deep learning approach to antibiotic discovery. *Cell*, *180*(4), 688–702.

Stone, T., & Choi, S.-K. (2013). *Extracting Consumer Preference From User-Generated Content Sources Using Classification* (Vol. Volume 3A: 39th Design Automation Conference) [V03AT03A031].

Strauss, D., & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American statistical association*, *85*(409), 204–212.

Sudman, S. (1966). Probability sampling with quotas. *Journal of the American Statistical Association*, *61*(315), 749–771.

Susilo, W. H. (2016). An Impact of Behavioral Segmentation to Increase Consumer Loyalty: Empirical Study in Higher Education of Postgraduate Institutions at Jakarta. *Procedia - Social and Behavioral Sciences*, *229*, 183–195.

Tovares, N., Cagan, J., & Boatwright, P. (2013). Capturing Consumer Preference Through Experiential Conjoint Analysis. *ASME Paper No. DETC2013-12549*.

Train, K. (1986). *Qualitative choice analysis: Theory, econometrics, and an application to automobile demand* (Vol. 10). MIT press.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Tuarob, S., & Tucker, C. S. (2015). Quantifying product favorability and extracting notable product features using large scale social media data. *Journal of Computing and Information Science in Engineering*, *15*(3).

Tucker, C. S., & Kim, H. M. (2008). Optimal product portfolio formulation by merging predictive data mining with multilevel optimization. *Journal of Mechanical Design*, *130*(4).

Tucker, C. S., & Kim, H. M. (2009). Data-driven decision tree classification for product portfolio design optimization. *Journal of Computing and Information Science in Engineering*, *9*(4).

Tucker, C. S., & Kim, H. M. (2011). Trend mining for predictive product design. *Journal of Mechanical Design*, *133*(11).

Van Horn, D., Olewnik, A., & Lewis, K. (2012). Design analytics: Capturing, understanding, and meeting customer needs using big data. *ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 863–875.

Wang, C., Butts, C. T., Hipp, J. R., Jose, R., & Lakon, C. M. (2016). Multiple imputation for missing edge data: A predictive evaluation method with application to add health. *Social Networks*, *45*, 89–98.

Wang, J., Huang, P., Zhao, H., Zhang, Z., Zhao, B., & Lee, D. L. (2018). Billion-scale commodity embedding for e-commerce recommendation in alibaba. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 839–848.

Wang, J., Chiu, K., & Fuge, M. (2020). Learning to abstract and compose mechanical device function and behavior. *ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*.

Wang, L., Youn, B. D., Azarm, S., & Kannan, P. K. (2011). Customer-driven product design selection using web based user-generated content. *ASME 2011 international design engineering technical conferences and computers and information in engineering conference*, 405–419.

Wang, M., Huang, Y., Contractor, N., Fu, Y., Chen, W., et al. (2016). A network approach for understanding and analyzing product co-consideration relations in engineering design. *DS 84: Proceedings of the DESIGN 2016 14th International Design Conference*, 1965–1976.

Wang, M., & Chen, W. (2015). A Data-Driven Network Analysis Approach to Predicting Customer Choice Sets for Choice Modeling in Engineering Design. *Journal of Mechanical Design*, *137*(7), 071410.

Wang, M., Chen, W., Fu, Y., & Yang, Y. (2015). Analyzing and Predicting Heterogeneous Customer Preferences in China's Auto Market Using Choice Modeling and Network Analysis. *SAE International Journal of Materials and Manufacturing*, *8*(2015-01-0468), 668–677.

Wang, M., Chen, W., Huang, Y., Contractor, N. S., & Fu, Y. (2015). A Multidimensional Network Approach for Modeling Customer-Product Relations in Engineering Design. *ASME 2015 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*.

Wang, M., Chen, W., Huang, Y., Contractor, N. S., & Fu, Y. (2016). Modeling customer preferences using multidimensional network analysis in engineering design. *Design Science*, *2*.

Wang, M., Sha, Z., Huang, Y., Contractor, N., Fu, Y., & Chen, W. (2018). Predicting product co-consideration and market competitions for technology-driven product design: A network-based approach. *Design Science*, *4*.

Wang, P., Robins, G., Pattison, P., & Lazega, E. (2013). Exponential random graph models for multilevel networks. *Social Networks*, *35*(1), 96–115.

Wang, X., He, X., Cao, Y., Liu, M., & Chua, T.-S. (2019). Kgat: Knowledge graph attention network for recommendation. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 950–958.

Wang, Z., Kannan, P., & Azarm, S. (2011). Customer driven optimal design for convergence products. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, *54822*, 379–394.

Wassenaar, H. J., & Chen, W. (2003). An approach to decision-based design with discrete choice analysis for demand modeling. *J. Mech. Des.*, *125*(3), 490–497.

Wassenaar, H. J., Chen, W., Cheng, J., & Sudjianto, A. (2005). Enhancing discrete choice demand modeling for decision-based design. *Journal of Mechanical Design*, *127*(4), 514–523.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp. *Psychometrika*, *61*(3), 401–425.

Williams, H., & Ortúzar, J. d. D. (1982). Behavioural theories of dispersion and the mis-specification of travel demand models. *Transportation Research Part B: Methodological*, *16*(3), 167–219.

Windzio, M. (2018). The network of global migration 1990–2013: Using ergms to test theories of migration between countries. *Social Networks*, *53*, 20–29.

Witten, I. H., & Frank, E. (2002). Data mining: Practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record*, *31*(1), 76–77.

Wu, L., Chen, Y., Shen, K., Guo, X., Gao, H., Li, S., Pei, J., Long, B., et al. (2023). Graph neural networks for natural language processing: A survey. *Foundations and Trends® in Machine Learning*, *16*(2), 119–328.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*.

Xie, J., Bi, Y., Sha, Z., Wang, M., Fu, Y., Contractor, N., Gong, L., & Chen, W. (2020). Data-driven dynamic network modeling for analyzing the evolution of product competitions. *Journal of Mechanical Design*, *142*(3).

Yin, F., & Butts, C. T. (2020). Kernel-based approximate bayesian inference for exponential family random graph models. *arXiv preprint arXiv:2004.08064*.

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., & Leskovec, J. (2018). Graph convolutional neural networks for web-scale recommender systems. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 974–983.

Zhang, T., Gensler, S., & Garcia, R. (2011). A Study of the Diffusion of Alternative Fuel Vehicles: An Agent-Based Modeling Approach. *Journal of Product Innovation Management*, *28*(2), 152–168.

Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2018). Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*.

# MULTI-STAGE CUSTOMER PREFERENCES MODELING USING DATA-DRIVEN NETWORK ANALYSIS

Approved by:

Wei Chen
Mechanical Engineering
*Northwestern University*

Noshir Contractor
Industrial Engineering and Management Science
*Northwestern University*

Elizabeth Gerber
Mechanical Engineering
*Northwestern University*

Date Approved: July 30, 2023

# APPENDIX A
# SURVEY QUESTIONNAIRE ON VACUUM CLEANER MARKET

## A.1 Filtering Questions

1. What kind of household products did you purchase within the past 12 months?

   A. Washing Machine

   B. Vacuum cleaner

   C. Flat-screen TV

   D. Refrigerator

   E. None of the above

2. Did you discuss vacuum cleaner options with anyone before your purchase?

   A. Yes

   B. No

## A.2 Consent to Participate in a Research Study

## A.3 Welcome to the Vacuum Cleaner Survey!

This survey is part of a research project being conducted by research groups at Northwestern University and the University of Arkansas. The purpose of this research project is to analyze customers' purchase preferences and identify key factors that influence customers' decision-making behaviors during online shopping. The product considered in this survey is the household vacuum cleaner. You are invited to participate in this web-based online survey because you purchased a vacuum cleaner within the last 12 months.

This survey is divided into the following 6 main parts:

1. Models considered and final purchase: This part is related to your purchase decision-making process. We ask which vacuum cleaner models you considered purchasing and which vacuum cleaner model you ultimately purchased.

2. Social network influence questions: This part asks you to list the individuals whose viewpoints were important to you in the past 12 months and who influenced your vacuum cleaner purchase decisions. The questions ask about the nature of your social relationships with these individuals and their demographic information.

3. Factors influencing decision-making: This part asks you to recall the factors that influenced which models you considered as well as the model you purchased.

4. Personal viewpoint: This part includes descriptive questions about your general purchase preferences for household appliances.

5. Using your vacuum cleaner: This part includes questions related to the use of your vacuum cleaner.

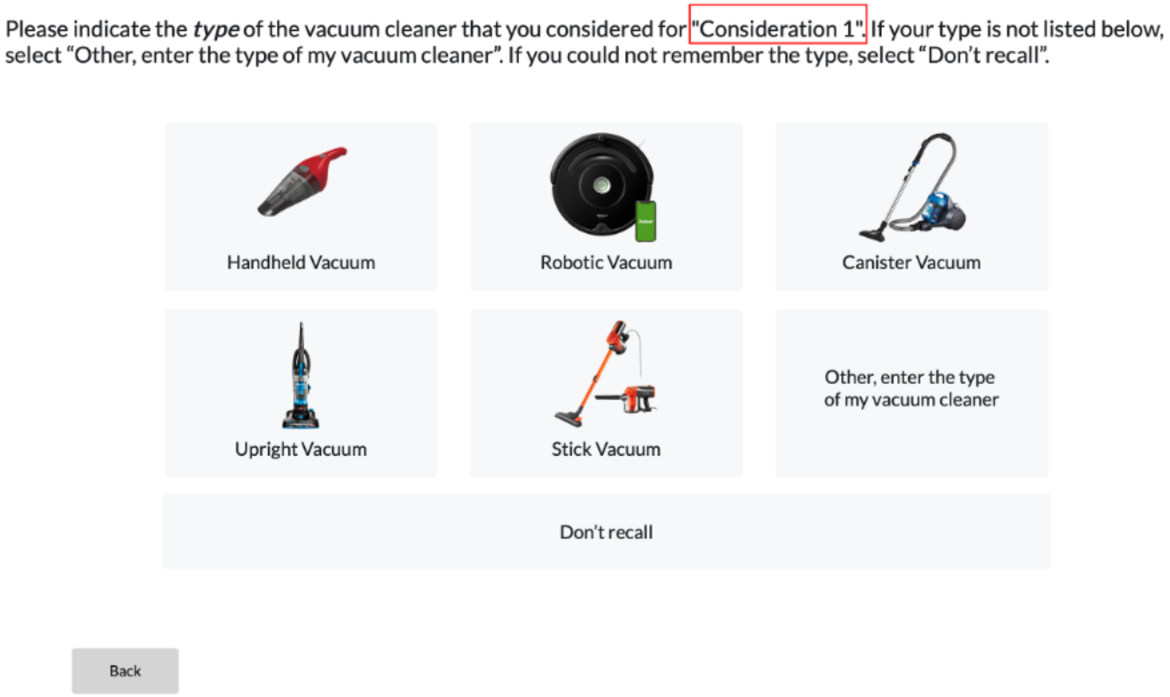
6. Demographic information: This part includes some demographic-related questions.

Note: Questions marked with an asterisk(*) are compulsory. It is important that you read and answer all required questions in this survey. As a check to test whether careful attention is being paid to each question, in some sections, you will be directed to select a specific answer.

Estimated time to complete: 30 minutes

## A.4 Purchase Decision-Making Process

(Shown on the survey website, this part can refer to the pilot study 1)



Please indicate the ***type*** of the vacuum cleaner that you considered for "Consideration 1". If your type is not listed below, select "Other, enter the type of my vacuum cleaner". If you could not remember the type, select "Don't recall".

## A.5 Your Social Network

### A.5.1 Discussion Details

1. Before we learn more about your vacuum cleaner purchase, we would like to learn something about your social network. People sometimes share and discuss important issues with others.

   (a) Looking back over the last 12 months, with whom did you discuss important issues with most often? Please list the names of at least 1 and up to 5 such individuals. Please do not use full names; first names only, initials, or nicknames are acceptable. (Example relationships: Spouse, Parent, Sibling, Child, Family member, Co-worker, Neighbor, Friend, Advisor, Stranger, Acquaintance, Other (dropdown))

(b) How often do you talk with each of these individuals? (Talking Frequency: Every day, Once a week, Once a month, Once every three months, Almost never)

(c) How many times did you talk about your vacuum cleaner purchase with these individuals within one month prior to the purchase? (Talking Frequency: 1-2 times, 3-4 times, More than 4 times, Never)

(d) In addition to the individuals identified above, if you also discussed the purchase of your vacuum cleaner with other people (for example, a salesperson), please list their names below and indicate your relationships with them. Please do not use full names; first names only, initials, or nicknames are acceptable. (Example relationships: Spouse, Parent, Sibling, Child, Other family member, Co-worker, Neighbor, Friend, Advisor, Stranger, Acquaintance, Salesperson, Other (dropdown))

(e) How often do you talk with each of these individuals? (subsequent question) (Talking Frequency: Every day, Once a week, Once a month, Once every three months, Almost never)

(f) How many times did you talk about your vacuum cleaner purchase with these individuals within one month prior to the purchase? (subsequent question)

### A.5.2 Demographic Details

2. We would now like to ask you some questions about the individuals you just identified.

   (a) Gender:

   - Female

- Male
- Non-binary
- Do not know
- Prefer not to say

(b) Age (in years):

- Under 18
- 18 - 24
- 25 - 34
- 35 - 44
- 45 - 54
- 55 - 64
- 65 - 74
- 75 - 84
- 85 or older
- Do not know
- Prefer not to say

(c) Ethnicity:

- African American
- Asian
- Caucasian
- Latino or Hispanic
- Native Hawaiian or Pacific Islander

- Other
- Do not know
- Prefer not to say

(d) Marriage status:

- Married
- Not married
- Do not know
- Prefer not to say

(e) Final level of full-time education or training:

- Grade School
- High school
- Trade school
- Community college
- Bachelor’s/4-yr degree
- Postgraduate degree
- Other
- Do not know
- Prefer not to say

(f) Occupation:

- Armed services
- Business Professional
- Clerical

- Craftsperson, Precision production
- Driver
- Entry-level professional
- Fabricator, Laborer
- Healthcare Professional
- Mid-level manager
- Owner, Self-employed
- Police, Fire, EMT
- Programmer, IT
- Sales
- Senior executive
- Service worker (food, cleaning)
- Specialty Worker
- Student
- Teacher, Educator
- Technician
- Stay-at-home parent/Homemaker
- Unemployed
- Retired
- Other
- Do not know
- Prefer not to say

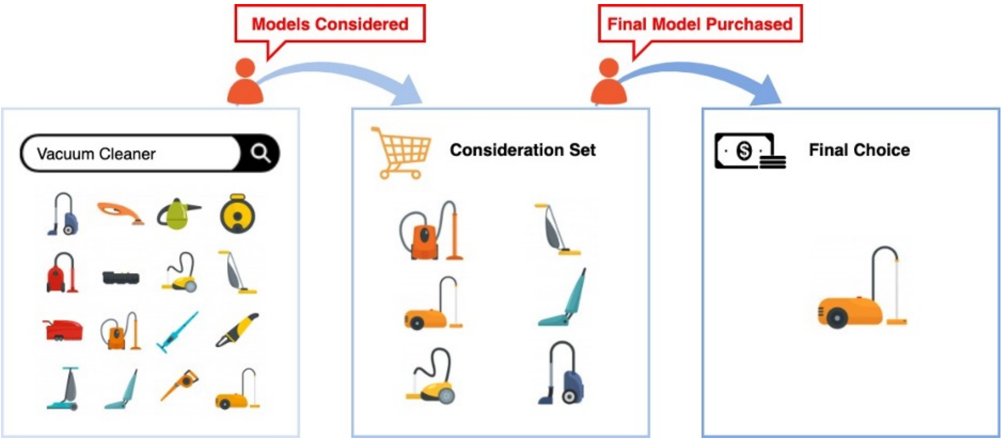
(g) Annual income:

- $10,000 or less
- $10,001 - $40,000
- $40,001 - $70,000
- $70,001 - $100,000
- $100,001 - $130,000
- $130,001 - $160,000
- Over $160,000
- Do not know
- Prefer not to say

Do you know if these individuals own the same vacuum cleaner you do?

- Yes, same make and model
- Yes, same make but different model
- Yes, same type but different make and model
- Neither same make nor model
- This individual does not have one
- I don’t know

Please indicate whether the people you mentioned know each other. Add any of the names to the boxes by dragging them in. Note that if person A and person B know each other, please drag A to the box of B or drag B to the box of A. You do not need to do both. If a person does not know any of the others, you do not need to drag their name into any box, or any name into their box.



## A.6 Factors Influencing Decision-making

Please note that this section contains two subsections: ”Models Considered” and ”Final Model Purchased”. Both subsections have identical questions but refer to different aspects of your vacuum cleaner purchase. Please ensure you are in the correct subsection before responding to the questions.

In the ”Models Considered” subsection, we ask about the products you considered before making the final decision.

In the ”Final Model Purchased” subsection, we ask the same set of questions but they pertain to the product you actually bought.

### A.6.1 Models Considered

- Please recall your recent vacuum cleaner purchase experience, and then evaluate how important each of the following sources of information was in influencing your decisions about which vacuum cleaners to consider. (Use a scale of 1 to 5 where 1 is ”not at all important” and 5 is ”very important”.)

- How important were each of the following individuals in influencing your decisions about which vacuum cleaners to consider? (Use the same scale of 1 to 5 as mentioned above.)

- Please recall your recent vacuum cleaner purchase experience, then drag and rank at least three features from the following list that most influenced your decision about which vacuum cleaners you considered. If you believe several features are equally important, please drag them into the same box. However, to proceed, you should fill at least the top 3 ranks (boxes). If the feature is not listed, please type in the feature.

### A.6.2 Final Model Purchased

- Reflect on your recent vacuum cleaner purchase experience, and then evaluate how important each of the following sources of information was in influencing your decision about which vacuum cleaner to purchase. (Use a scale of 1 to 5 where 1 is ”not at all important” and 5 is ”very important”.)

- How important were each of the following individuals in influencing your decision about which vacuum cleaner to purchase? (Use the same scale of 1 to 5 as mentioned above.)

- Please recall your recent vacuum cleaner purchase experience, then drag and rank at least three features from the following list that most influenced your decision about which vacuum cleaner to purchase. If you believe several features are equally important, please drag them into the same box. However, to proceed, you should fill at least the top 3 ranks (boxes). If the feature is not listed, please type in the feature.

### A.6.3 Purchase Satisfaction Questions

Reflecting on your experience with the model you purchased, please drag and drop the features of your purchased vacuum cleaner with which you are satisfied and dissatisfied into the corresponding columns in the table below. If you are neither satisfied nor dissatisfied with any of the features, select the ”None of the above” option.

## A.7 Views about Vacuum Cleaners

Please tell us how much you agree or disagree with the following statements.

1 - strongly disagree 2 - somewhat disagree 3 - neutral 4 - somewhat disagree 5 - strongly disagree

1. As far as vacuum cleaners are concerned, I am always looking for an innovative model.

2. When buying a vacuum cleaner I like it if manufacturers add the most modern technology to it.

3. The type of vacuum cleaner I buy needs to reflect my lifestyle.

4. I would pay more for environmental-friendly features.

5. Styling is at or near the top of the important characteristics in a new vacuum cleaner.

6. I only buy vacuum cleaners with good energy efficiency.

7. When I buy a vacuum cleaner I choose the least expensive one that meets my needs.

8. I will buy the vacuum cleaner that is the easiest to maintain.

9. Please choose Neutral.

10. I keep a vacuum cleaner as long as possible.

11. Exceptional after-sales service is the reason enough to warrant buying the same make time after time.

12. I always buy the same make of vacuum cleaner.

13. When deciding which make of vacuum cleaner to buy, I take seriously what other people have to say.

14. I consider myself an advocate of my favorite vacuum cleaner brand, telling others about my experience.

15. I would pay more for the highest quality vacuum cleaner.

## A.8 Using Your Vacuum Cleaner

1. What is your current state of residence?

   State:

2. Which city do you currently reside in?

   City:

3. What is your current living arrangement?

   - A home you own
   - A home owned by your parent(s) or other family member(s)
   - A home owned by a friend
   - A rental home

4. What type of home do you live in?

    - Single house
    - Townhouse
    - Apartment
    - Condo
    - Other

5. Does your home have stairs?

    - Yes
    - No

6. How many rooms are there in your home (including living rooms, bedrooms, kitchens, and bathrooms)?

    - 1 - 5
    - 6 - 10
    - 11 - 15
    - Over 15

7. What types of flooring are in your home? Please check all that apply.

    - Carpet
    - Wooden
    - Tile

- Vinyl

8. Which number below is the largest one?

    - 21
    - 4
    - 23
    - 15

9. How many pets with fur or hair (e.g., cats/dogs) live in your home?

    - 0
    - 1
    - 2
    - 3
    - Over 3

10. How often do you cook?

    - Every day
    - 1-2 days a week
    - 3-4 days a week
    - 5-6 days a week
    - Never

11. How often is your vacuum cleaner used in your home?

- Every day
- Every week
- Every month
- Every year
- Only in response to a spill or mess
- Never

12. Do you have a cleaning service that cleans your home?

    - Yes
    - No

13. How often does the cleaning service clean your home? (Answer only if question 12 is ”Yes”.)

    - Every day
    - Every week
    - Every month
    - Every year

14. Does the cleaning service use your vacuum cleaner or bring their own? (Answer only if question 12 is ”Yes”.)

    - Uses my vacuum cleaner
    - Brings their own

### A.9 Demographic Attributes

1. What gender do you identify as?

   - Female
   - Male
   - Non-binary
   - Prefer not to say

2. Please indicate your age (in years).

   - Under 18
   - 18 - 24
   - 25 - 34
   - 35 - 44
   - 45 - 54
   - 55 - 64
   - 65 - 74
   - 75 - 84
   - 85 or older
   - Prefer not to say

3. Please indicate your ethnicity.

   - African American

- Asian
- Caucasian
- Latino or Hispanic
- Native Hawaiian or Pacific Islander
- Other
- Prefer not to say

4. Please indicate your marital status.

   - Married
   - Not married
   - Prefer not to say

5. What is the highest degree or level of education you have completed?

   - Grade school
   - High school
   - Trade school
   - Community college
   - Bachelor’s/4-yr degree
   - Postgraduate degree
   - Other

6. Which of the following best describes your occupation?

- Armed services
- Business Professional
- Clerical
- Craftsperson, Precision production
- Driver
- Entry-level professional
- Fabricator, Laborer
- Healthcare Professional
- Mid-level manager
- Owner, Self-employed
- Police, Fire, EMT
- Programmer, IT
- Sales
- Senior executive
- Service worker (food, cleaning)
- Specialty Worker
- Student
- Teacher, Educator
- Technician
- Stay-at-home parent/Homemaker
- Unemployed

- Retired
- Other

7. What’s your annual household income?

    - $10,000 or less
    - $10,001 - $40,000
    - $40,001 - $70,000
    - $70,001 - $100,000
    - $100,001 - $130,000
    - $130,001 - $160,000
    - Over $160,000
    - Prefer not to say

8. Including yourself, how many people normally live in your household in total?

    - Only me
    - 2-3
    - 4-5
    - 6-7
    - Over 7

9. Of those living in your household, other than you, how many are males, females and/or non-binary? Please enter ”0” when a category is empty.

- Female:
- Male:
- Non-binary:

10. How many children aged under 18 years normally live in your household in total?

- None
- 1
- 2
- 3
- Over 3

# APPENDIX B

# SURVEY RESULTS COMPREHENSIVE VIEW ON CUSTOMER FEATURES

We provide summary statistics of the survey data to help the audience get an overview of the dataset.

Figure B.1 represents the customers' two-stage decision-making process, displaying a histogram of the number of vacuum cleaners (other than the purchase one) all respondents considered.

Figure B.2 shows a histogram of the number of people in each respondent's social network.

Figure B.3 displays a count plot of the factors influencing customers' consideration and purchase stages based on respondents' ranking in the survey data. The plot shows a weighted sum of the respondents' rankings, assigning higher weights to features that received higher rankings. Therefore, the plot presents the weighted feature importance ranking.

Figures B.4, B.5, B.6 depict histograms of personal viewpoints about vacuum cleaners, usage context questions, and demographic questions in the survey, respectively.

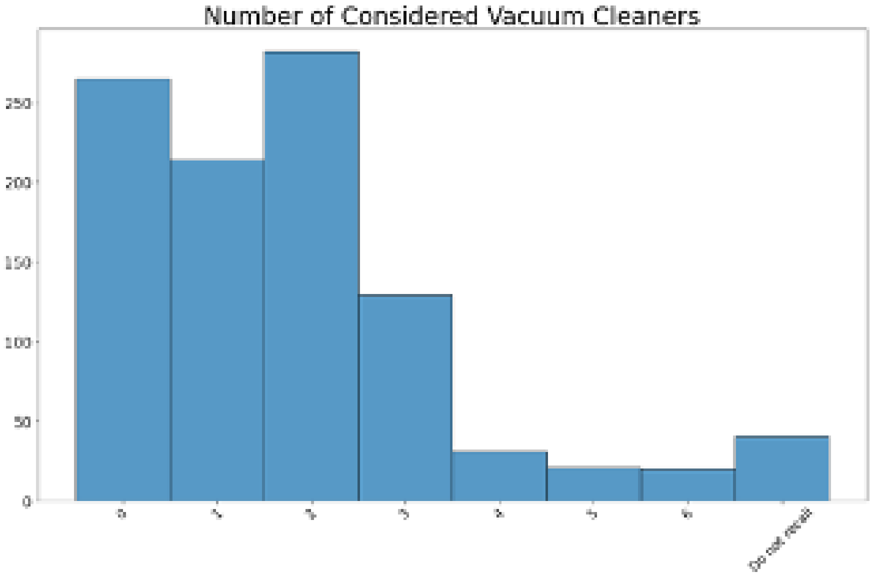


Figure S B.1: Histogram of the number of vacuum cleaners (other than the purchase one) considered by respondents.

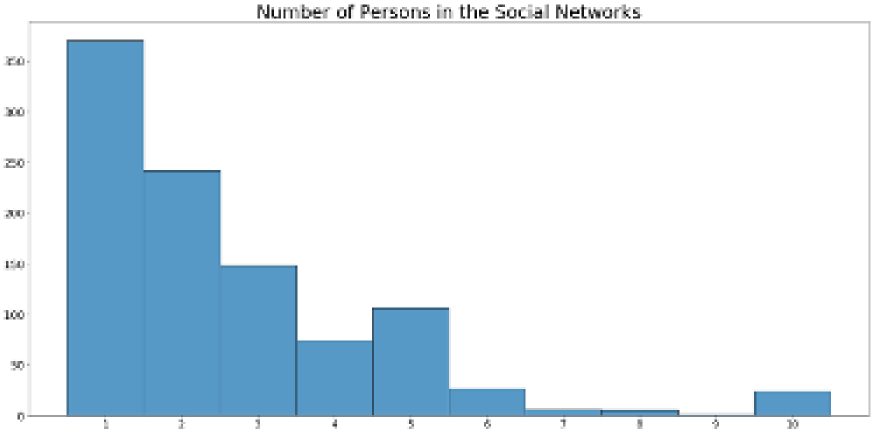


Figure S B.2: Histogram of the number of people in respondents' social networks.

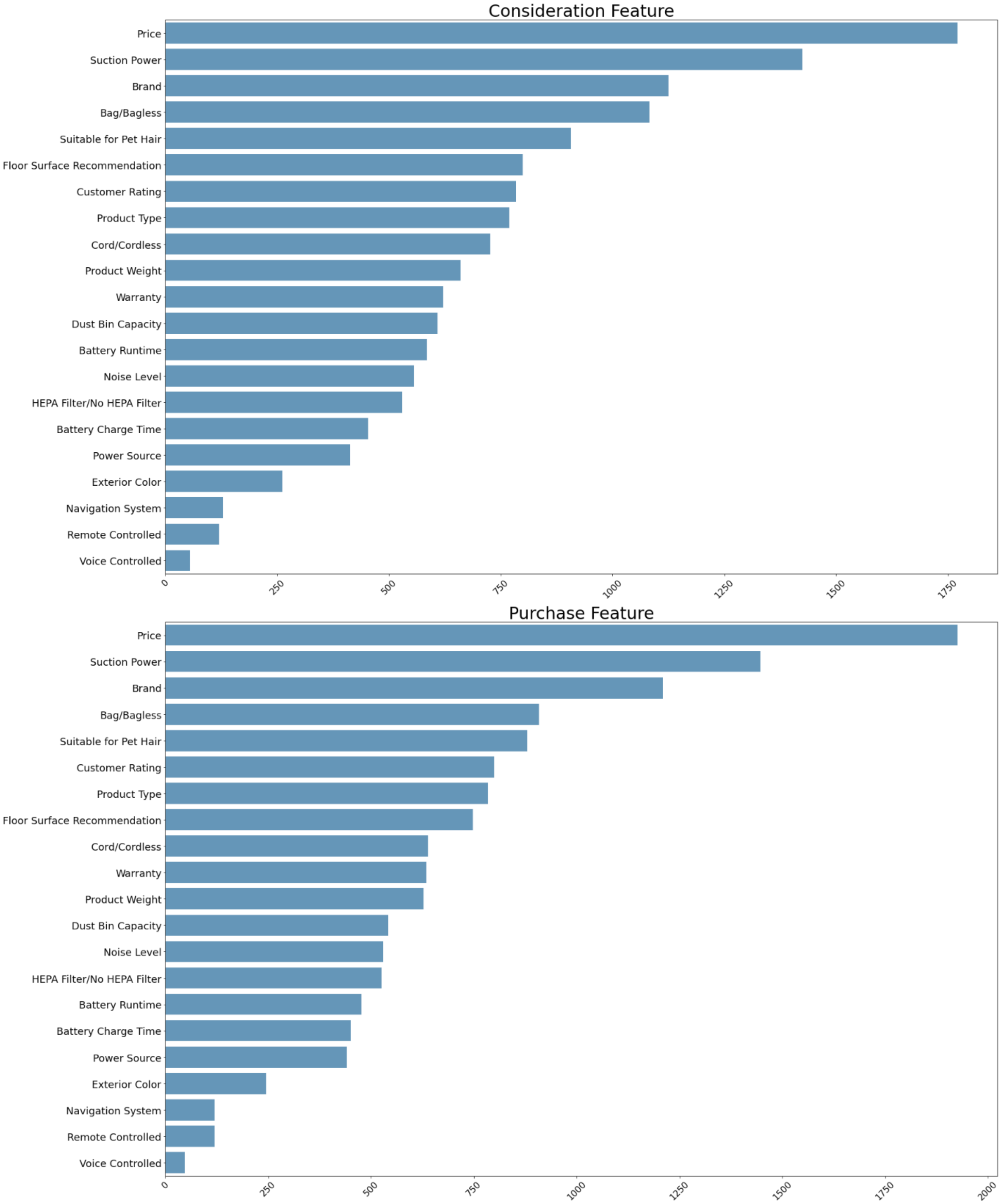


Figure S B.3: Weighted feature importance rankings reported by respondents for consideration and purchase (choice) stages.

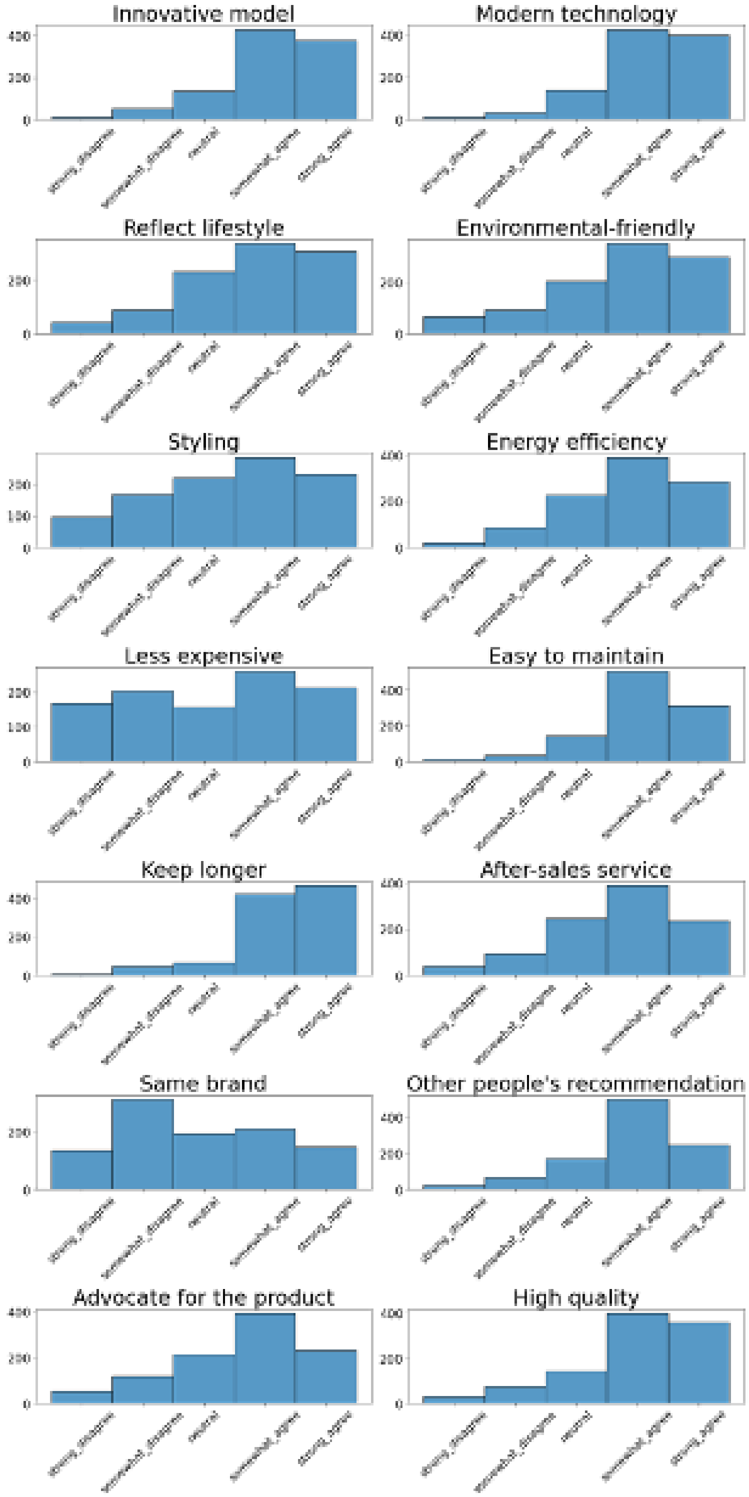


Figure S B.4: Histogram of variables related to respondents' personal viewpoints about vacuum cleaners.

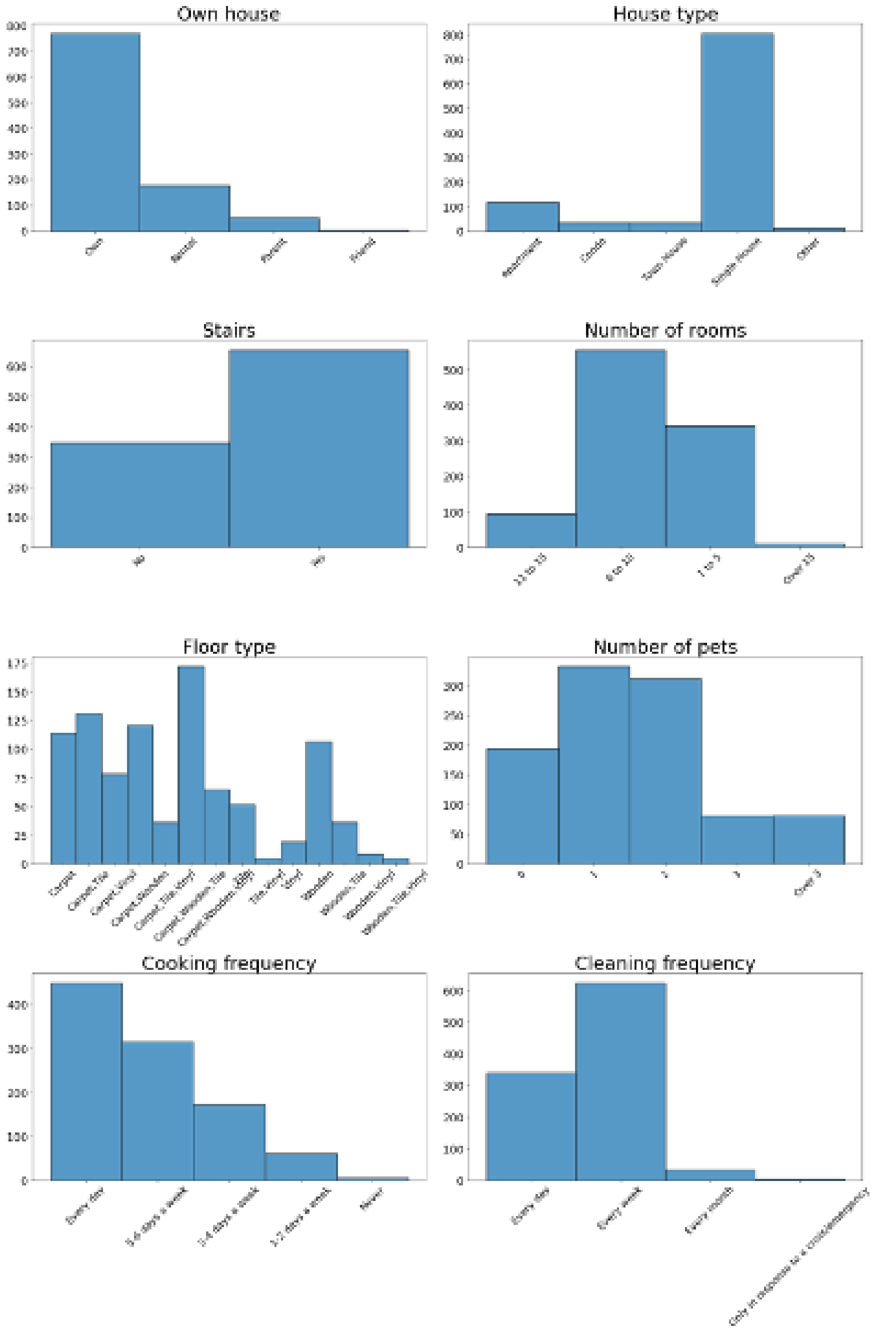


Figure S B.5: Histogram of variables related to respondents’ usage context questions.

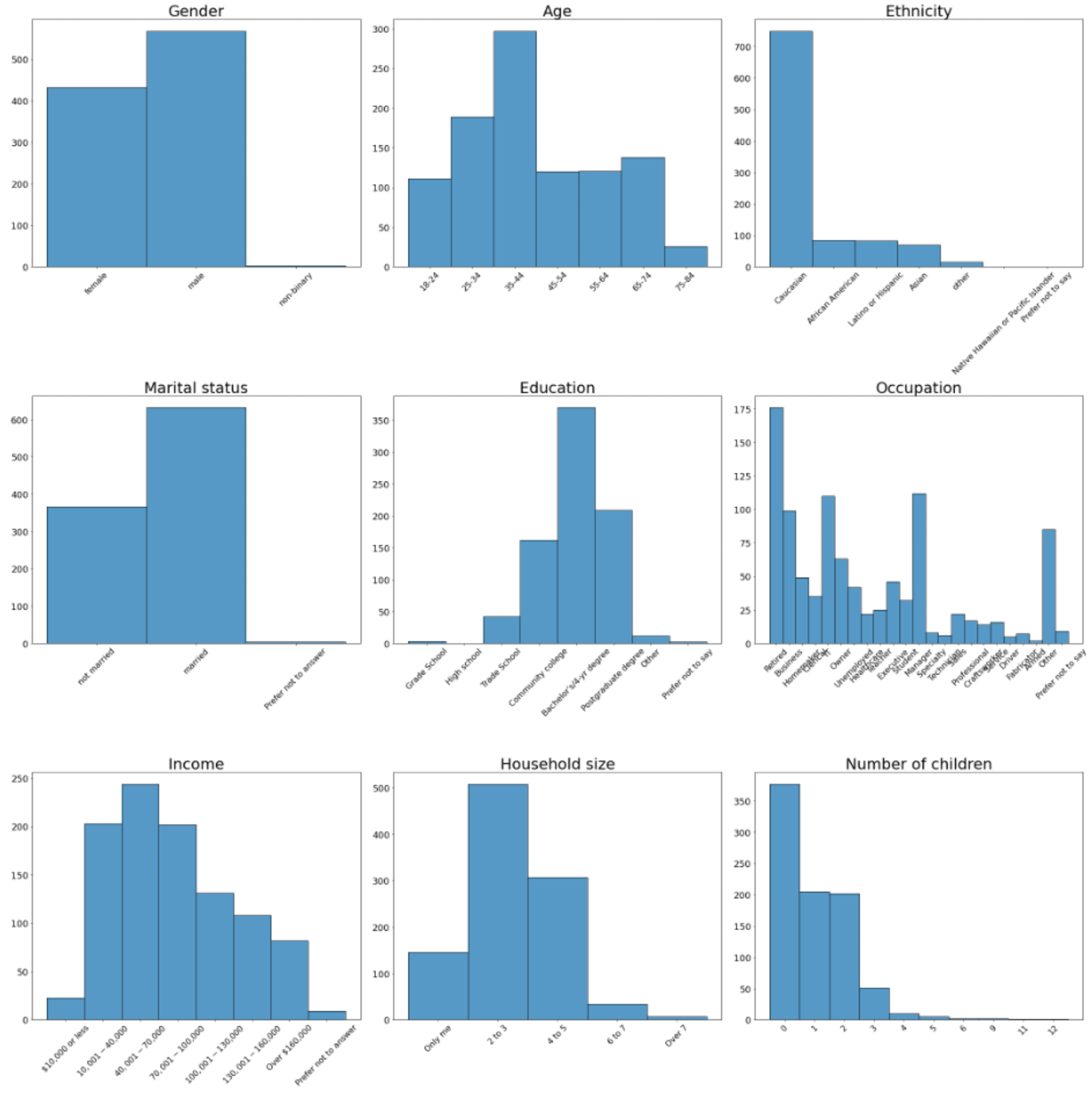


Figure S B.6: Histogram of variables related to respondents' demographic questions.

# VITA

## NAME

Yaxin Cui

## PLACE OF BIRTH

Shanxi, P.R. CHINA

## EDUCATION

NORTHWESTERN UNIVERSITY, Evanston, Illinois, USA    Sept. 2018 - June. 2023

**Ph.D.** in Mechanical Engineering

SHANGHAI JIAO TONG UNIVERSITY, Shanghai, China    Sept. 2014 - Jul. 2018

**B.S.** in Mechanical Engineering

## EMPLOYMENT

FORD MOTOR COMPANY    Jun. 2021-Sept. 2021

**Summer Research Intern at Global Data Analytics and Insight**

SAIC VOLKSWAGEN AUTOMOTIVE COMPANY    Mar. 2018-Jun. 2018

**Bachelor Thesis Intern for Handling Electric Battery Pack**

## TEACHING EXPERIENCES

NORTHWESTERN UNIVERSITY

| | |
|---|---|
| **Grader, Computational Methods for Engineering Design Class** | Winter 2021 & 2022 |
| **Grader, Fluid Mechanics Class** | Spring 2021 |

## HONORS AND AWARDS

NORTHWESTERN UNIVERSITY

| | |
|---|---|
| **Travel Grant Award** | Oct. 2022 |
| **Intersection Science Fellowship by Data Science Institute** | Dec. 2020 |
| **Design Fellowship by Segal Design Cluster** | Sep. 2019 |
| **Walter P. Murphy Fellowship** | Sept. 2018 |
| **International Summer Institute Scholarship** | Aug. 2018 |

SHANGHAI JIAO TONG UNIVERSITY

| | |
|---|---|
| **Outstanding Graduate in Shanghai** | June. 2018 |
| **Honorable Mention in MCM/ICM** | Apr. 2017 |
| **Fan Hsu-chi Scholarship** | Oct. 2016 & 2017 |
| **Guanghua Scholarship** | Nov. 2015 |

## JOURNAL ARTICLES

**Cui, Y.**, Sun, Z., Xiao, Y., Sha, Z., Koskinen, J., Contractor, N., & Chen, W. (In preparation). *Network-Based Analysis of Heterogeneous Consideration-then-Choice Customer Preferences with Market Segmentation.* Journal of Interactive Marketing, 2023.

Xiao, Y., **Cui, Y. (co-first author)**, Raut, N., Januar, H., Koskinen, J., Contractor, N., Chen, W.,

& Sha, Z. (under review). *Survey Data on Customer Two-Stage Decision-Making Process in Household Vacuum Cleaner Market.* Data in Brief, 2023.

Sha, Z., **Cui, Y.**, Xiao, Y., Stathopoulos, A., Noshir, C, Fu, Y., & Chen, W. (2023). *A network-based discrete choice model for decision-based design.* Design Science, 9, E7.

**Cui, Y.**, Ahmed, F., Sha, Z., Wang, L., Fu, Y., & Chen, W. (2022). *A Weighted Statistical Network Modeling Approach to Product Competition Analysis.* Complexity.

Ahmed, F., **Cui, Y.**, Fu, Y., & Chen, W. (2022). *Product Competition Prediction in Engineering Design using Graph Neural Networks.* ASME Open J. Engineering ASME.

## BOOK CHAPTERS

Chen, W., Ahmed, F., **Cui, Y.**, Sha, Z., & Contractor, N. (2022). *Data-Driven Preference Modelling in Engineering Systems Design.* In *Systems Engineering Handbook,* eds. Maier, Anja, Oehmen, Josef, Vermaas, Pieter E., Springer.

## REFERRED CONFERENCE PAPERS (PEERED REVIEWED)

**Y. Cui**, Y. Xiao, Z. Sha, W. Chen (2023). *Network-Based Analysis of Heterogeneous Consideration-then Choice Customer Preferences with Market Segmentations.* In *2023 Conference on Systems Engineering Research (CSER),* Hoboken, New Jersey. Mar. 16-17, 2023. (In press)

Y. Xiao, **Y. Cui**, W. Chen, Z. Sha (2023). *Product Competition Analysis for Engineering Design: A Network Mining Approach.* In *2023 Conference on Systems Engineering Research (CSER),* Hoboken, New Jersey. Mar. 16-17, 2023. (In press)

Malhotra, P., **Cui, Y.**, & Zhao, K. (2022). *Modeling Co-Engagement Patterns in Brand Information Networks.* In *2022 IEEE 16th International Conference on Semantic Computing*

*(ICSC),* Laguna Hills, CA, USA, pp. 257-262.

Ahmed, F., **Cui, Y.**, Fu, Y., & Chen, W. (2021, August). *A graph neural network approach for product relationship prediction.* In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference,* American Society of Mechanical Engineers.

**Cui, Y.**, Ahmed, F., Sha, Z., Wang, L., Fu, Y., & Chen, W. (2020, August). *A weighted network modeling approach for analyzing product competition.* In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference,* American Society of Mechanical Engineers.

## INVITED PRESENTATIONS

- ”Network-Based Analysis of Heterogeneous Consideration-then Choice Customer Preferences with Market Segmentations,” In 2023 Conference on Systems Engineering Research (CSER), Hoboken, New Jersey. Mar. 16-17, 2023.

- “Network-based Customer Preference Modeling,” NICO Lightening Talk at Northwestern University, Oct 19, 2022, Evanston, IL, USA

- “Network-based Customer Preference Modeling,” Lambert Conference on the Future of Human-Computer Interaction + Design, Oct 24-25, 2022, Evanston, IL, Poster.

- ”Modeling Co-Engagement Patterns in Brand Information Networks,” INFORMS Annual Meeting, Oct 16-19, 2022, Indiana, USA

- “Network-based Customer Preference Modeling,” Sunbelt 2022 – The XLII International Sunbelt Social Networks Conference, July 12-16, 2022, Cairns, Australia. (Presented online)

- ”A weighted network modeling approach for analyzing product competition,” International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, 2020, Online