Design Science

A network-based discrete choice model for decision-based design

Zhenghui Sha[1], Yaxin Cui[2], Yinshuang Xiao[1], Amanda Stathopoulos[3], Noshir Contractor[4], Yan Fu[5] and Wei Chen[2]

[1] *Walker Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX, USA*
[2] *Department of Mechanical Engineering, Northwestern University, Evanston, IL, USA*
[3] *Department of Civil and Environmental Engineering, Northwestern University, Evanston, IL, USA*
[4] *Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL, USA*
[5] *Global Data Insight and Analytics, Ford Motor Company, Dearborn, MI, USA*

Abstract

Customer preference modelling has been widely used to aid engineering design decisions on the selection and configuration of design attributes. Recently, network analysis approaches, such as the exponential random graph model (ERGM), have been increasingly used in this field. While the ERGM-based approach has the new capability of modelling the effects of interactions and interdependencies (e.g., social relationships among customers) on customers' decisions via network structures (e.g., using triangles to model peer influence), existing research can only model customers' consideration decisions, and it cannot predict individual customer's choices, as what the traditional utility-based discrete choice models (DCMs) do. However, the ability to make choice predictions is essential to predicting market demand, which forms the basis of decision-based design (DBD). This paper fills this gap by developing a novel ERGM-based approach for choice prediction. This is the first time that a network-based model can explicitly compute the probability of an alternative being chosen from a choice set. Using a large-scale customer-revealed choice database, this research studies the customer preferences estimated from the ERGM-based choice models with and without network structures and evaluates their predictive performance of market demand, benchmarking the multinomial logit (MNL) model, a traditional DCM. The results show that the proposed ERGM-based choice modelling achieves higher accuracy in predicting both individual choice behaviours and market share ranking than the MNL model, which is mathematically equivalent to ERGM when no network structures are included. The insights obtained from this study further extend the DBD framework by allowing explicit modelling of interactions among entities (i.e., customers and products) using network representations.

Keywords: customer preference modelling, exponential random graph model, multinomial logit model, decision-based design

1. Introduction

As a vital link between market research and engineering design, discrete choice models (DCMs) predict a customer's choice probability through the construction of a utility function that quantitatively characterises the customer's preferences for product design attributes (Chen *et al.* 2018). Choice models have been used to

Received 27 August 2022
Revised 15 February 2023
Accepted 16 February 2023

Corresponding author:
Z. Sha
zsha@austin.utexas.edu

Des. Sci., vol. 9, e7
journals.cambridge.org/dsj
DOI: 10.1017/dsj.2023.4

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

support many aspects of engineering design and are the foundation of the decision-based design (DBD) framework (Chen, Hoyle & Wassenaar 2012). The applications include conceptual design (Hoyle & Chen 2009), multidisciplinary design (MacDonald, Gonzalez & Papalambros 2009), product configuration (Sha *et al.* 2017*a*), product innovation (Chang & Chen 2014; Chen, Khoo & Chen 2015) and design accounting for spatiotemporal heterogeneities (Bi *et al.* 2018). In a design ecosystem that includes multiple stakeholders (e.g., manufacturers and policy-makers), understanding customer preferences is crucial in collaboration among stakeholders with varied interests and allowing them to make strategic planning (Kang *et al.* 2016; Chen *et al.* 2020).

Customer preference and choice modelling in many engineering design applications with large numbers of alternatives and decision-makers can be addressed by network modelling. As shown in Figure 1, the customer–product relationship can be represented by a bipartite network, where customers and products are modelled as two types of nodes, and the considerations and choices of the customers are modelled as different types of links. Therefore, decision analysis in this context can be viewed as modelling the likelihood of links forming between nodes. Taking the example of vehicle choice analysis, typically the large set of available vehicles and their buyers can be represented as a network graph, and link prediction represents the problem of estimating events (choices) connecting customers to vehicle options by edges. Therefore, formulating the decision analysis in a network context means that we are interested in understanding what factors (either exogenous or endogenous) drive the formation of a link (choice/consideration) between a pair of a customer node and a product node, and how significant a role each factor plays in that link formation process.

Across network modelling approaches, there is a common theme of relational analysis where decisions (events) connect nodes in a network. Logit models from the discrete choice family naturally model such relational events as ‘choices’ (McFadden & Zarembka 1974; Luce 2012). In the past two decades, utility-based choice modelling, such as DCMs (Train 1986), has been widely employed by the engineering design research community (Frischknecht, Whitefoot & Papalambros 2010; Hoyle *et al.* 2010; He *et al.* 2014) for choice preference estimation and demand prediction. Nevertheless, utility-based choice modelling is limited when handling the *dependency of alternatives* (e.g., it is assumed that whether a customer chooses one product is not influenced by adding or substituting another product in the choice set, which is not realistic for applications with similar product offerings) and the *rationality of customer decisions* (e.g., their decisions may be influenced by each other due to social relationships). To overcome these limitations, recent studies explored the capability of statistical network models in customer choice analysis (Fu *et al.* 2017; Sha *et al.* 2018). Among the existing network-based modelling techniques, the exponential random graph model (ERGM) is increasingly recognised as one of the most powerful analytical techniques (Snijders *et al.* 2006). ERGM provides a flexible statistical inference framework that can model the influence of both exogenous effects (e.g., nodal attributes that include design features or customer attributes) and endogenous effects (network structures/nodal relations) on the probability of connections among nodes.

ERGM can handle three types of networks based on the complexity of the customer–product network: unidimensional, bipartite and multidimensional networks, as shown in Table 1. In our previous work, ERGM has been used to analyse

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

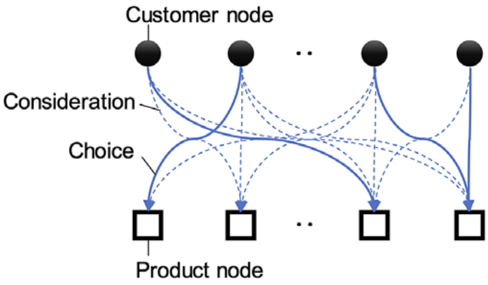
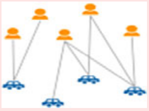


Figure 1. The customer–product bipartite choice network.

Table 1. Networks with different complexities

Network type	Illustration	Description	Reference papers
Unidimensional		A unidimensional network with only product nodes can describe the product competition relationship based on aggregated customer preferences.	Wang *et al.* (2016*b*), Sha *et al.* (2017*b*), Sha *et al.* (2018), Wang *et al.* (2018), Xie *et al.* (2020), Ahmed *et al.* (2021), Cui *et al.* (2022)
Bipartite		A bipartite network defines the relation (consideration or choices) between the customer layer and the product layer but not within each individual layer.	Fu *et al.* (2017), Bi *et al.* (2018, 2021)
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.	Wang *et al.* (2015, 2016*a*)

and predict product co-consideration relations (Sha *et al.* 2018; Wang *et al.* 2018; Xie *et al.* 2020; Cui *et al.* 2022), forecast the impact of technological changes on market competition (Wang *et al.* 2016*b*), model customers' consideration-then-choice behaviours (Fu *et al.* 2017) and understand how product associations and customer social relationships affect customer decisions (Wang *et al.* 2016*a*).

In these studies, attempts were made to examine the predictive power of ERGM approaches. For example, ERGM was used to predict product co-consideration relations (implying competition relations) in the scenario of product design upgrades and market changes (Wang *et al.* 2016*b*; Sha *et al.* 2018). Additionally, temporal ERGM was adopted to predict market evolution based on historical data. However,

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

all existing predictive studies focus on unidimensional networks. To the best of our knowledge, no studies were carried out to investigate the power of ERGM in predicting individual customer's choices in a bipartite network setting. This gap exists because of the lack of a mathematical framework for ERGM to calculate the probability of individual dyad links in a network. In ERGM, the random variable in such a statistical inference framework is the network (see Section 2 for more details). Therefore, one ERGM prediction corresponds to one entire network. To test the model's predictive performance on individual links, the current procedure is to simulate the random process in network generation to predict many networks sharing statistical similarities with the true network data, from which the correctly predicted links can be counted. This limitation has led to two fundamental issues when adopting ERGM in choice modelling. First, the concept of 'choice set' cannot be applied. Second, the concept of 'utility' in traditional DCMs cannot be imported. Without these two concepts, a formal choice analysis cannot be formulated. Therefore, a correct interpretation of customer preferences could never be reached when ERGM is applied for choice modelling, and the strength of ERGM to incorporate network structures in choice prediction cannot be fully extended for engineering design. To exploit the full potential of ERGM in choice modelling and, therefore, the design-based design, a new mathematical procedure for ERGM-based choice prediction must be developed and a new approach that applies such a procedure in support of engineering design must be created.

Motivated by filling the gap, we aim to answer two research questions in this paper. One: What is the mathematical procedure that enables ERGM to compute individual choice probabilities in a given choice set? Two: Given that ERGM can successfully predict choices, do the prediction results differ from applying traditional DCM (i.e., the baseline model) to the same decision scenario and dataset? And, if the results are different, how can these differences be interpreted? The answers to these questions will contribute to an improved understanding of network-based approaches in modelling customer preferences for design.

The contributions of this paper can be summarised in three aspects: (1) The development of a new mathematical procedure, enabling ERGM to compute the probability of an individual link, and, therefore, an individual customer's choices. (2) A systematic approach of applying ERGM to conduct formal choice analysis taking into account the concept of choice set. This helps ERGM to correctly predict product demand, which is critical to implementing the DBD framework. (3) A theoretical insight into the differences between the utility-based and network-based approaches in discrete choice analysis that could lead to valuable guidance in choosing an appropriate model for customer preference modelling and co-developing the two frameworks in support of engineering design and beyond.

The remainder of the paper is structured as follows. In Section 2, the background knowledge about ERGM and its three major categories of network statistics are reviewed. The proposed ERGM-based analytical framework of single dyad link prediction is introduced in Section 3. The corresponding model settings and the evaluation methods for predictive power are also presented in this section. Section 4 presents the case study using a vehicle market system as an example, including the introduction of the data source, the choice scenarios and the discussion on the model estimation and prediction results. Section 5 discusses the implications of the new ERGM-based choice model and how it can support and extend the DBD framework, which is followed by a comparison of DCMs and

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

ERGMs in a design application. The paper is concluded in Section 6 with a presentation of the limitations of this study and future work.

2. Exponential random graph model

The ERGMs are a family of statistical inference models for network data analysis (Harris 2013). The ERGM defines a probability model of an observed network $\mathbf{y}$, as one specific realisation from a set of possible random networks Y, following the distribution in the following equation:

$$\Pr(\mathbf{Y}=\mathbf{y})=\frac{\exp(\boldsymbol{\theta}'\mathbf{g}(\mathbf{y}))}{\kappa(\boldsymbol{\theta})}, \tag{1}$$

where $\boldsymbol{\theta}$ is a vector of model parameters, $\mathbf{g}(\mathbf{y})$ is a vector of the network statistics defining various network structures that can incorporate either nodal attributes or edge attributes and $\kappa(\boldsymbol{\theta})$ is a normalising quantity to ensure that Eq. (1) is a proper probability distribution. Eq. (1) suggests that the probability mass function on the network space is proportional to the exponential of a linear combination of network statistics. The formulation also indicates that the network with statistic in $\mathbf{g}(y)$ is more likely to occur if the corresponding $\boldsymbol{\theta}$ is positive.

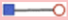
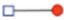
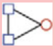
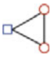
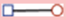
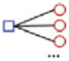
The strength of ERGM is its capability of modelling endogenous interdependencies (i.e., relations) among entities (e.g., products) with various forms of network statistics, that is, $\mathbf{g}(\mathbf{y})$, in addition to exogenous attributes pertaining to nodes and/or edges. Typically, the network statistics can be categorised into three main categories, that is, nodal attribute effects, relational attribute effects and network structural effects (Morris, Handcock & Hunter 2008), as shown in Table 2. Nodal attribute effects refer to the main effects of the nodes, which can be either continuous covariates or discrete (e.g., categorical) variables. In a customer–vehicle network, nodal attributes could be car features (e.g., price and engine size) and customer-related attributes (e.g., income and education level). Relational attribute effects measure the effects of dyads' (i.e., a group of two nodes) and edges' attributes. Examples of relational attributes include the similarity of dyad attributes and edge covariates. Moreover, network structural effects measure the different levels of complexity of network structures, including the basic terms that control the overall probability of a link (such as the number of edges and density of a network), degree and star attributes which capture the distribution of node-based edge counts, and triangles and higher-order cycles that measure the effect of more complex local network structures.

3. Methodology

In this section, we first present a stepwise ERGM-based choice modelling approach (see Figure 2) that guides choice analysis, estimation, prediction and evaluation. The approach consists of five steps. In the first step, the revealed product choices are identified, and the choice set is defined along with the alternative-specific attributes obtained from a large-scale product database. Then, the customer–product bipartite networks are constructed and the choice scenarios are defined. Step 2 is about ERGM modelling that includes identifying critical network statistics and utilising additional settings and constraints (see Section 3.2) to ensure that the ERGM-based choice models accurately capture the given choice scenarios. Based

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

Table 2. Three major categories of network statistics in exponential random graph models in a customer–product network

Category	Examples	Interpreted effects
Nodal attributes effects		Car attributes (price and fuel consumption)
		Customer attributes (income and education level)
Relational attributes effects		Customer considers similar products
		Peer influence among customers
Network structural effects		Density
		Star effects and product popularity

Note: The blue square represents the product, and the red circle represents the customer. The solid shape refers to the node of interest in modelling instead of the dyad relation.

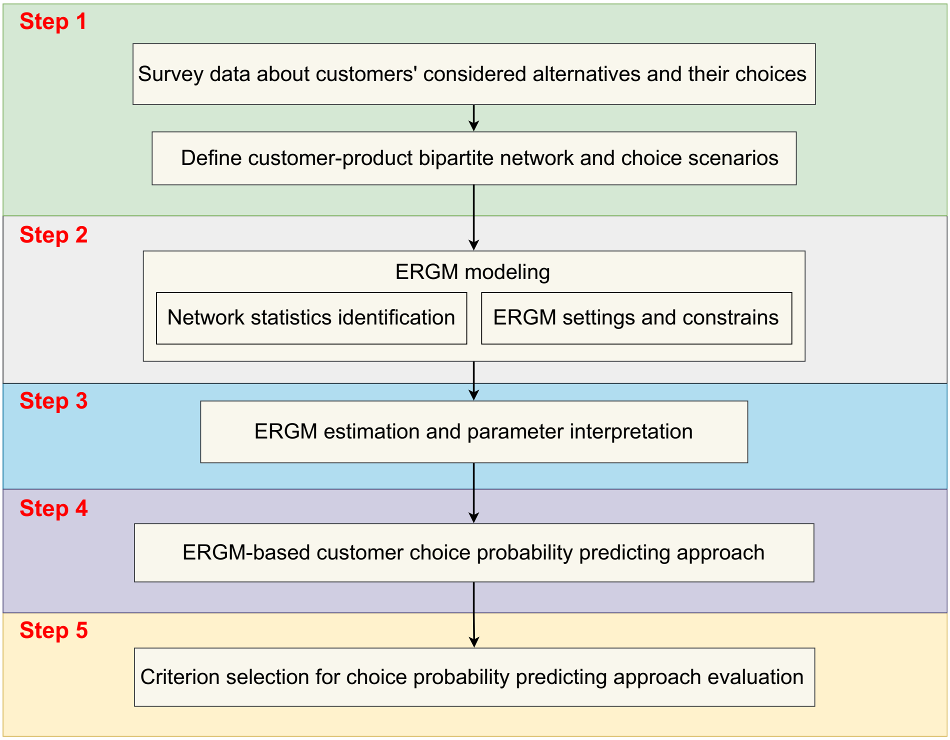

Step 1
Survey data about customers' considered alternatives and their choices
Define customer-product bipartite network and choice scenarios
Step 2
ERGM modeling
Network statistics identification
ERGM settings and constrains
Step 3
ERGM estimation and parameter interpretation
Step 4
ERGM-based customer choice probability predicting approach
Step 5
Criterion selection for choice probability predicting approach evaluation

Figure 2. Overview of the exponential random graph model-based choice modelling research approach.

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

on the models established in Step 2, Step 3 estimates the model parameters and compares the results of the models with different configurations (see Section 4.2). This step is critical because, by analysing similarity and differences, we build a knowledge base to understand consistency and interpretability of distinct ERGMs when applied to the same choice problem. In Step 4, the proposed ERGM-based choice prediction method is implemented. The mathematical procedure and the derivation are presented in Section 3.1. Step 5 is to select appropriate criteria to evaluate the predictive performance of the ERGM-based choice models. The criteria adopted in this study are based on two metrics: the Top-*N* probability (see Section 3.3.1) and the market share ranking (see Section 3.3.2). It should be noted that Steps 4 and 5 are essential to help draw valuable insights on the distinctions and connections between traditional utility-based and network-based approaches in choice modelling.

3.1 A new mathematical procedure for computing the predicted choice probability using ERGM

In this section, we propose a novel method for calculating the choice probability (i.e., the linking probability in the network context) based on ERGM. To predict the existence of a particular link, an important concept, called *change statistics*, must be introduced. *Change statistics* emerge when considering the probability of a single dyad having a link given the rest of the network. The vector of *change statistics* of a link is defined as follows (Hunter *et al.* 2008):

$$\Delta_{ij}\mathbf{g}(\mathbf{y}) \equiv \mathbf{g}(\mathbf{y}_{\mathbf{ij}}^{+}) - \mathbf{g}(\mathbf{y}_{\mathbf{ij}}^{-}), \tag{2}$$

where $\mathbf{y}_{\mathbf{ij}}^{+}$ is the network $\mathbf{y}$ with edge (i,j) added if absent (and unchanged if present), and $\mathbf{y}_{\mathbf{ij}}^{-}$ is the network $\mathbf{y}$ with edge (i,j) removed if present (and unchanged if absent). Then, based on Eq. (1), the probabilities of both networks $\mathbf{y}_{\mathbf{ij}}^{+}$ and $\mathbf{y}_{\mathbf{ij}}^{-}$ can be expressed as follows:

$$\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(\mathbf{Y}=\mathbf{y}_{\mathbf{ij}}^{+}) = \frac{\exp(\boldsymbol{\theta}'\mathbf{g}(\mathbf{y}_{\mathbf{ij}}^{+}))}{\kappa(\boldsymbol{\theta})}, \tag{3}$$

$$\mathrm{Pr}_{\theta;\mathbf{g}}(\mathbf{Y}=\mathbf{y}_{\mathbf{ij}}^{-}) = \frac{\exp(\theta'\mathbf{g}(\mathbf{y}_{\mathbf{ij}}^{-}))}{\kappa(\theta)}, \tag{4}$$

By dividing those two formulas, we obtain

$$\frac{\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(\mathbf{Y}=\mathbf{y}_{\mathbf{ij}}^{+})}{\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(\mathbf{Y}=\mathbf{y}_{\mathbf{ij}}^{-})} = \exp(\boldsymbol{\theta}'\mathbf{g}(\mathbf{y}_{\mathbf{ij}}^{+}) - \mathbf{g}(\mathbf{y}_{\mathbf{ij}}^{-})) = \exp(\boldsymbol{\theta}\cdot\Delta_{ij}\mathbf{g}(\mathbf{y})). \tag{5}$$

We can further write the formulas of conditional probabilities:

$$\frac{\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(Y_{ij}=1|\mathbf{y}^{\mathbf{c}})}{\mathrm{Pr}_{\theta;\mathbf{g}}(Y_{ij}=0|\mathbf{y}^{\mathbf{c}})} = \frac{\mathrm{Pr}_{\theta;\mathbf{g}}(\mathbf{Y}=\mathbf{y}_{\mathbf{ij}}^{+})/\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(\mathbf{Y}_{\mathbf{c}}=\mathbf{y}_{\mathbf{c}})}{\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(\mathbf{Y}=\mathbf{y}_{\mathbf{ij}}^{-})/\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(\mathbf{Y}_{\mathbf{c}}=\mathbf{y}_{\mathbf{c}})} = \exp(\boldsymbol{\theta}\cdot\Delta_{ij}\mathbf{g}(\mathbf{y})), \tag{6}$$

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

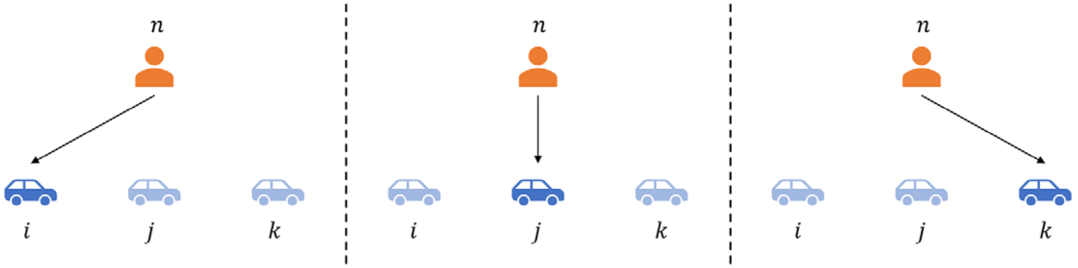


Figure 3. Illustration of the choice situation where a customer makes a choice among three alternatives. A customer n is allowed to make a purchase decision among the three car models, car i, car j and car k. The figure illustrates the three choice scenarios.

Given the fact that $\Pr_{\boldsymbol{\theta};\mathbf{g}}(Y_{ij}=1|\mathbf{y}^{\mathbf{c}})+\Pr_{\boldsymbol{\theta};\mathbf{g}}(Y_{ij}=0|\mathbf{y}^{\mathbf{c}})=1$, we can rewrite the equation as follows:

$$\Pr_{\boldsymbol{\theta};\mathbf{g}}(Y_{ij}=1|\mathbf{y}^{\mathbf{c}})=\operatorname{logit}^{-1}(\boldsymbol{\theta}\cdot\Delta_{ij}\mathbf{g}(\mathbf{y})), \tag{7}$$

where $\mathbf{y}^{\mathbf{c}}$ is the remaining network structure without the edge (i,j), and $\Pr_{\boldsymbol{\theta};\mathbf{g}}(Y_{ij}=1|\mathbf{y}^{\mathbf{c}})$ denotes how likely the link $Y_{i,j}$ is to exist given the remaining network structure. Eq. (7) presents the mathematical foundation of the network link prediction (i.e., the choice prediction) based on the ERGM formula together with the *change statistics.*

In the multinominal choice scenario, where each customer can have only one choice among multiple alternatives, the formula of link prediction in a bipartite customer–product network can be developed. Without loss of generality, we assume that each customer has three products (e.g., cars) in a choice set, and a customer n is allowed to make a single purchase decision among the three car models, i, j and k (as illustrated in Figure 3).

Based on the assumption that a customer chooses only one among the three alternatives, the following relation of the link probability holds:

$$\begin{aligned}&\Pr_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=1,Y_{nj}=0,Y_{nk}=0|\mathbf{y}^{\mathbf{c}})+\\&\Pr_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=0,Y_{nj}=1,Y_{nk}=0|\mathbf{y}^{\mathbf{c}})+\\&\Pr_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=0,Y_{nj}=0,Y_{nk}=1|\mathbf{y}^{\mathbf{c}})=1,\end{aligned} \tag{8}$$

where Y_{ni},Y_{nj},Y_{nk} denote the link between customer n and each car model, respectively, and $\mathbf{y}^{\mathbf{c}}$ denote the remaining network structure.

The conditional probability of link existence between node pairs (n,i),(n,j) and (n,k) can be expressed as follows:

$$\Pr_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=1,Y_{nj}=0,Y_{nk}=0|\mathbf{y}^{\mathbf{c}})=\frac{\Pr_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=1,Y_{nj}=0,Y_{nk}=0,\mathbf{Y}^{\mathbf{c}}=\mathbf{y}^{\mathbf{c}})}{\Pr_{\boldsymbol{\theta};\mathbf{g}}(\mathbf{Y}^{\mathbf{c}}=\mathbf{y}^{\mathbf{c}})}, \tag{9}$$

$$\Pr_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=0,Y_{nj}=1,Y_{nk}=0|\mathbf{y}^{\mathbf{c}})=\frac{\Pr_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=0,Y_{nj}=1,Y_{nk}=0,\mathbf{Y}^{\mathbf{c}}=\mathbf{y}^{\mathbf{c}})}{\Pr_{\boldsymbol{\theta};\mathbf{g}}(\mathbf{Y}^{\mathbf{c}}=\mathbf{y}^{\mathbf{c}})}, \tag{10}$$

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

$$\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=0,Y_{nj}=0,Y_{nk}=1|\mathbf{y^c})=\frac{\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=0,Y_{nj}=0,Y_{nk}=1,\mathbf{Y^c}=\mathbf{y^c})}{\mathrm{Pr}_{\theta;\mathbf{g}}(\mathbf{Y^c}=\mathbf{y^c})}. \tag{11}$$

Further, the marginal probability of numerators in Eqs. (9)–(11) can be represented by the ERGM in Eq. (1). Thus, we can rewrite Eqs. (9)–(11) as follows:

$$\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=1,Y_{nj}=0,Y_{nk}=0|\mathbf{y^c})=\frac{\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni}^+,\mathbf{nj}^-,\mathbf{nk}^-}))}{\kappa(\boldsymbol{\theta},\mathbf{y})\cdot\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(\mathbf{Y^c}=\mathbf{y^c})}, \tag{12}$$

$$\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=0,Y_{nj}=1,Y_{nk}=0|\mathbf{y^c})=\frac{\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni}^-,\mathbf{nj}^+,\mathbf{nk}^-}))}{\kappa(\boldsymbol{\theta},\mathbf{y})\cdot\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(\mathbf{Y^c}=\mathbf{y^c})}, \tag{13}$$

$$\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=0,Y_{nj}=0,Y_{nk}=1|\mathbf{y^c})=\frac{\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni}^-,\mathbf{nj}^-,\mathbf{nk}^+}))}{\kappa(\boldsymbol{\theta},\mathbf{y})\cdot\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(\mathbf{Y^c}=\mathbf{y^c})}, \tag{14}$$

where $\mathbf{y}_{\mathbf{ni}^+,\mathbf{nj}^-,\mathbf{nk}^-}$ denotes the network $\mathbf{y}$ with link $Y_{n,i}$ present and links $Y_{n,j}$ and $Y_{n,k}$ absent, $\mathbf{y}_{\mathbf{ni}^-,\mathbf{nj}^+,\mathbf{nk}^-}$ denotes the network $\mathbf{y}$ with link $Y_{n,j}$ present and links $Y_{n,i}$ and $Y_{n,k}$ absent, $\mathbf{y}_{\mathbf{ni}^-,\mathbf{nj}^-,\mathbf{nk}^+}$ denotes the network $\mathbf{y}$ with link $Y_{n,k}$ present and links $Y_{n,i}$ and $Y_{n,j}$ absent.

By taking the division of Eqs. (12) and (13),

$$\frac{\mathrm{Pr}_{\theta;\mathbf{g}}(Y_{ni}=1,Y_{nj}=0,Y_{nk}=0|\mathbf{y^c})}{\mathrm{Pr}_{\theta;\mathbf{g}}(Y_{ni}=0,Y_{nj}=1,Y_{nk}=0|\mathbf{y^c})}=\exp[\boldsymbol{\theta}\cdot(\mathbf{g}(\mathbf{y}_{\mathbf{ni}^+,\mathbf{nj}^-,\mathbf{nk}^-})-\mathbf{g}(\mathbf{y}_{\mathbf{ni}^-,\mathbf{nj}^+,\mathbf{nk}^-}))]. \tag{15}$$

By taking the division of Eqs. (12) and (14),

$$\frac{\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=1,Y_{nj}=0,Y_{nk}=0|\mathbf{y^c})}{\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=0,Y_{nj}=0,Y_{nk}=1|\mathbf{y^c})}=\exp[\boldsymbol{\theta}\cdot(\mathbf{g}(\mathbf{y}_{\mathbf{ni}^+,\mathbf{nj}^-,\mathbf{nk}^-})-\mathbf{g}(\mathbf{y}_{\mathbf{ni}^-,\mathbf{nj}^-,\mathbf{nk}^+}))]. \tag{16}$$

Given a linear system of three equations of (8), (15) and (16), we can further solve for three unknown conditional probabilities as

$$\begin{aligned}&\mathrm{Pr}_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=1,Y_{nj}=0,Y_{nk}=0|\mathbf{y^c})=\\&\frac{\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni}^+,\mathbf{nj}^-,\mathbf{nk}^-}))}{\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni}^+,\mathbf{nj}^-,\mathbf{nk}^-}))+\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni}^-,\mathbf{nj}^+,\mathbf{nk}^-}))+\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni}^-,\mathbf{nj}^-,\mathbf{nk}^+}))}.\end{aligned} \tag{17}$$

This is exactly the probability that the link Y_{ni} exists given the rest of network structures, that is, the customer n chooses car model i given the rest of customer choices. Similarly, the probability that a customer chooses the other two products, j and k, can be expressed in the following equations:

$$\Pr_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=0, Y_{nj}=1, Y_{nk}=0|\mathbf{y^c}) = \frac{\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni^-},\mathbf{nj^+},\mathbf{nk^-}}))}{\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni^+},\mathbf{nj^-},\mathbf{nk^-}}))+\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni^-},\mathbf{nj^+},\mathbf{nk^-}}))+\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni^-},\mathbf{nj^-},\mathbf{nk^+}}))}. \tag{18}$$

$$\Pr_{\boldsymbol{\theta};\mathbf{g}}(Y_{ni}=0, Y_{nj}=0, Y_{nk}=1|\mathbf{y^c}) = \frac{\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni^-},\mathbf{nj^-},\mathbf{nk^+}}))}{\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni^+},\mathbf{nj^-},\mathbf{nk^-}}))+\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni^-},\mathbf{nj^+},\mathbf{nk^-}}))+\exp(\boldsymbol{\theta}\cdot\mathbf{g}(\mathbf{y}_{\mathbf{ni^-},\mathbf{nj^-},\mathbf{nk^+}}))}. \tag{19}$$

Eqs. (17)–(19) are the mathematical foundations of predicting customer choices based on the ERGM formula.

Based on the format of the probability distribution of link Y_{ni}, Y_{nj}, Y_{nk}, we note that if our network model only has exogenous effects (i.e., car attributes and customer attributes) as input to the model, the vector of the network statistic $\mathbf{g}(\mathbf{y})$ will be a linear combination of only car attributes and customer attributes. Then, the probability distribution in Eqs. (17)–(19) will be equivalent to that of the ordinal multinomial logit (MNL) model (McFadden *et al.* 1973) under the assumptions aforementioned. In other words, when the ERGM only considers the nodal attributes but not the network statistics, the choice prediction using the ERGM can degenerate to an ordinary MNL model if certain model constraints are properly included. See the following subsection for details. However, if endogenous effects (i.e., network statistics) are considered, changes in network structure will have an impact on the predicted linking probability. In this case, the choice prediction based on ERGM will be different from that of DCM. In this study, we specifically investigated how the inclusion of network statistics in the choice prediction process influences the predictive accuracy of ERGM-based choice models.

3.2 ERGM settings and constraints

Traditional DCA is centred on decision analysis and makes model assumptions that chosen variables and choice sets remain stable throughout the modelling of a given decision scenario. These assumptions, however, do not hold in ERGM. For example, by default, there is no choice set defined in ERGM. The application of ERGM for different scenarios is often realised by means of different model configurations, that is, different combinations of edge and node settings and constraints. To enable the ERGM to model a general purchase decision scenario, two network statistics (*b2factor* (Morris *et al.* 2008) and *b2cov* (Handcock & Hunter 2018)) and two ERGM constraints (*offset* (Hunter *et al.* 2008) and *b1degrees* (Handcock & Hunter 2018)) are adopted.

(i) *b2cov* and *b2factor*: The main effect of a covariate and factor attribute effect for the second mode in a bipartite network (Handcock & Hunter 2018). These two terms are used to carry quantitative and categorical nodal attributes

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

corresponding to the product features that can influence customer preferences and then as input variables $\mathbf{g}(\mathbf{y})$ for ERGM in Eq. (1).

(ii) *offset*: A vector of coefficients for the offset terms (Handcock & Hunter 2018). In the customer–product bipartite network, this constraint is used to control how many product nodes are considered to form the links. This represents how many products can be linked to (i.e., considered by) each customer. Without this constraint, the default setting of ERGM will assume that each customer could link with all listed products.

(iii) $constraints = b1degrees$: Together with the model terms in the formula and the reference measure, the constraints define the distribution of the networks being modelled (Handcock & Hunter 2018). The constraint of $b1degrees$ adopted in this study is used to limit the assumption that each customer node only links to one product node. This means that we assume that each customer can only make one choice from the choice set.

3.3 Evaluation criteria for prediction accuracy

Given that there is no established framework to evaluate ERGM predictions, we adopt two criteria that go beyond the evaluation of just the chosen alternative to consider the overall ranking. The first criterion evaluates the accuracy of the predicted customers' Top-N choices, and the second one evaluates the predicted market share based on the predicted choice probability. These two criteria are introduced in the following two subsections.

*Criterion 1: Top-*N *choice probability*

Suppose that there are L customers and M products, and every customer has a choice set, including h products. The probability that the i^{th} $(i = 1,2,\ldots,L)$ customer buys the j^{th} $(j = 1,2,\ldots,h)$ product is represented as P_{ij}. With the predicted choice probabilities of the top N $(N = 1,2,\ldots,h)$ alternatives in a choice set, we compare it against a customer's final choice. If the predicted choice falls in the Top-N alternative set, then this is counted as one accurate prediction instance (Cremonesi, Koren & Turrin 2010; Mottini *et al.* 2018). For example, in the case of $N = 2$, if the actual choice is one of the two alternatives with the highest predicted probability, it is counted as a correct prediction. The advantage of the Top-N metric is that it does not only capture customers' final choices (i.e., the Top 1 choice), but also evaluates the model's performance in capturing a customer's consideration behaviour. This metric provides a comprehensive evaluation of the model's predictive power over the entire choice set. Based on this definition, the predictive accuracy can be defined as

$$Acc_{topN} = \frac{L_{predict}}{L}, \tag{20}$$

where $L_{predict}$ is the number of customers whose choices are correctly predicted to fall into the Top-N alternatives. With different configurations of the N value, a trend curve can be obtained to delineate how the predictive accuracy would be influenced by the change of N.

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

Criterion 2: market share ranking

While the Top-N choice probability criterion captures a choice model's performance in predicting individual customers' choices, the predictive power of models at the aggregate level, that is, the market level, is typically more relevant to enterprises. This approach reflects both the positioning and the specific market share of a product, thereby providing an indication of a product's market competitiveness. The prediction-based market share can be divided into two steps. The first step is to predict the product that the i^{th} customer purchased based on the maximum choice probability in the choice set ($\max P_{ij}, i = 1, 2, \ldots, L; 0 \le j \le h$). In the second step, we count the number of customers who purchase each unique product to derive each product's market share reflecting all L customers' purchase behaviour.

After obtaining the predicted market shares of the products of interest, we can rank them based on their market shares and compare this ranking to the (real) reference market share. To quantify the predictive accuracy, two metrics are used. The first metric is focused on the ranking, and the *Spearman's rank correlation coefficient* (Schober, Boer & Schwarte 2018) is adopted to verify how well the rank of the most popular products (e.g., the top-K products ranked by the market share) predicted by the model matches the real data. The second metric compares the exact market share percentages, and computes the *root mean square error (RMSE)* (Chai & Draxler 2014) from comparing the predicted market shares to the actual one, and is calculated based on the following equation:

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(Real\_MarketShare_k - Predict\_MarketShare_k\right)^2}. \tag{21}$$

There are several motivations for adopting these validation measures based on market share. First, the rank correlation coefficient of the market share reflects the models' ability to forecast the popularity of the products in the overall market. Second, the RMSE value measures the accuracy of the market share predicted by the models compared to the actual market shares. Lastly, since both the market share ranking and its exact percentages are critical indicators to help enterprises assess a product's market performance, the combination of the rank correlation coefficient and RMSE draws a more complete picture of the overall performance of the models.

4. Case study

4.1 Data preprocessing and scheme of test scenarios

Data source

The dataset used in this study is drawn from a large-scale revealed choice new car buyer survey conducted in the Chinese vehicle market in 2013. This dataset includes about 50,000 respondents and nearly 400 unique vehicle models. From this survey, various customer-related and car-related attributes are recorded, such as respondents' gender and age, and vehicles' price, power and brand. The database also records the stated consideration of vehicle models, in addition to the recorded final purchase. A more detailed summary of this dataset can be found in our previous paper (Sha *et al.* 2019). In line with our previous work, we narrow down the analysis to a single type of vehicle model. By focusing on the compact sedan

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

standard vehicle segment, we seek to simplify the modelling and remove some sources of heterogeneity in decisions and choice sets. The analysis subset includes 18,054 compact sedan buyers, covering 84 unique compact sedan car models. Note that the stated considerations of alternatives are not limited to the compact sedan segment, although the final purchase was restricted to compact sedan purchases, making the total number of different car models covered in the analysis 281. Lastly, to improve the computational efficiency, 5000 customers are randomly sampled from the 18,054 compact sedan buyers to form the models' training dataset[3]. To test the sensitivity of results to the sampled data, we created three additional testing datasets of 5000 random customers (not overlapped with the training dataset) for analysis and prediction.

The treatment of choice sets

In this case study, two treatments of the choice set are considered. In the first treatment, we rely on stated information from customers to define a choice set of variable composition but identically sized (i.e., including the same number of alternatives). In the second treatment, it is assumed that each customer has the same fixed universal choice set consisting of all vehicle choice alternatives available on the market. Based on these assumptions, the two choice sets are defined as follows:

(i) *ChoiceSet6*: The first approach pursues behavioural realism by including stated consideration alternatives in a limited-size choice set. For every N customer, we include information from the original survey data where they list between one to three 'considered vehicle types'. While it is tempting to run models exclusively on this declared choice set, we note that the high degree of feature correlation inherent in enumerating close 'runner-up' options, leads to singularity and non-convergence issues in the DCA model. Therefore, we use insight from the literature (Hauser & Wernerfelt 1990) that customers typically consider three to six different car models in the buying process to build a more diverse choice set. In practice, we build a synthetic *ChoiceSet6* where we append randomly selected car models from the complete set of 281 to each customer's original stated consideration.

(ii) *ChoiceSetAll*: The second approach is less behavioural but more comprehensive. We assume that the N customers each have the same universal choice set consisting of 281 car models.

The scheme of model test

Based on the two choice set scenarios, *ChoiceSet6* or *ChoiceSetAll*, our case study includes two test cases, as summarised in Table 3: *Test Case 1* and *Test Case 2.* Because each customer can only make one choice in either choice scenario, $constraints = b1degrees$ needs to be applied in both test cases to constrain the number of links formed between a customer and his/her choice set to one. In addition, the *offset* needs to be used in *Test Case 1* to make sure the link is formed between the customer and his/her own choice set of six

[3] The validation of whether the sampled training dataset effectively represents the original dataset can be found in appendix A

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

Table 3. The scheme of comparison

	Test Case 1	Test Case 2
Customer number	5000	5000
Choice set	*ChoiceSet6*	*ChoiceSetAll*
ERGM setting and constraints	*offset*, *b2cov*, *constraints* = *b1degree*	*b2cov*, *constraints* = *b1degree*
Considered car attributes	*Price*, *Fuel Consumption*, *Power*, *Brand Origin*	

Abbreviation: ERGM, exponential random graph model base choice.

Table 4. Descriptive statistics of the 281 sedan car attributes

Continuous variables	
Average *Price*	218,163 RMB (31,617 USD)
Average *Fuel Consumption*	9.56 litre per 100 km
Average *Power*	147.29 brake horsepower (BHP)
Categorical variables: *Brand Origin*	
China	86
The US	29
Europe	88
Japan	56
Korea	22

alternatives. Otherwise, the ERGM by default treats all available car nodes as being able to be connected.

We select a few representative yet significant explanatory variables based on our prior work (Fu *et al.* 2017). These explanatory vehicle attributes are: *Price*, *Fuel Consumption*, *Power* and *Brand Origin*. For all 281 car models in the 2013 Chinese market, the descriptive statistics of these four attributes are listed in Table 4. Additionally, given the large variation exhibited in each attribute and their skewed distributions, the log transformation (base 2) is applied to the price and engine power variables to offset the effect of large outliers, and z-score normalisation is applied to address the unit influence of those continuous variables.

4.2 ERGM estimation results

The estimated parameters of the ERGMs in the two cases are shown in Tables 5 and 6, respectively. Table 5 summarises the results of the *ChoiceSet6* test case. For comparative purposes, we specify the same set of exogenous variables (i.e., car attributes) for all the ERGMs. Since the ERGM is uniquely positioned to handle endogenous factors (e.g., different network structures), we investigated two network effects, the degree effect and the star effect. The degree effect quantifies the influence of the number of a node's connections on the formation of new links to that node. Hence, the degree effect reflects the popularity of a node. The star effect

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

Table 5. Estimated results of ChoiceSet6 test case with the comparison of different exponential random graph model base choices

Explanatory variables	$ERGM_{Null}$	$ERGM_{Degree}$	$ERGM_{Star}$	$ERGM_{Both}$
Price	−3.21305***	−3.22619***	−3.05046***	−3.02861***
Fuel consumption	−0.49992***	−0.49958***	−0.03704	−0.03438
Power	2.02397***	2.02632***	1.59739***	1.59093***
Brand origin (the US)	1.85477***	1.85549***	1.23290***	1.22953***
Brand origin (Europe)	2.96869***	2.97428***	2.21507***	2.20565***
Brand origin (Japan)	1.27489***	1.27974***	1.22087***	1.21286***
Brand origin (Korea)	1.25118***	1.27480***	1.18760***	1.16191***
Network degree (25+)		−1.09881**		1.20514**
Network 2-star			0.00578***	0.00583***
Running time	31 minutes	85 minutes	88 minutes	165 minutes

Note: (1) The code *** indicates the 0.001 level of significance and ** indicates the 0.005 level of significance.
(2) Network degree (25+) represents the number of product nodes that have more than 25 connections to customer nodes.
(3) Network 2-star represents the number of star-like structure with two edges connected to the central products node.

Table 6. Estimated results of ChoiceSetAll test case with the comparison of different exponential random graph model base choices

Input variables	$ERGM_{Null}$	$ERGM_{Star}$
Price	−3.14765***	−0.34672***
Fuel consumption	−0.70885***	−0.00457***
Power	2.17598***	0.18375***
Brand origin (the US)	2.39916***	0.49637***
Brand origin (Europe)	3.30098***	0.38671
Brand origin (Japan)	1.21206***	0.23710
Brand origin (Korea)	1.40266***	0.33653***
Network 2-star		0.05227****
Network 3-star		−0.00037***
Running time	7 seconds	18 seconds

Note: The code *** indicates the 0.001 level of significance.

is a star-like structure with multiple edges connected to the central node. For example, 2-star represents the structure where there is one node connecting two links, that is, one vehicle model is purchased by two customers in our case. Please note that a specific star structure could be embedded. For example, for a node with three links, it has a degree of 3, three 2-stars and one 3-star. So, both effects capture the distribution of node-based edge counts and reflect the degree distribution information. However, the star effect could amplify the popularity of a node because the number of stars (s) that could be obtained from a node with n degrees

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

is C_d^s. Please note that $C_d^s \geq d$ and the increase of one more degree will make the number of stars $(n+1)$ times greater. Therefore, there are four ERGMs to be tested. $ERGM_{Null}$ only considers exogenous factors and does not include any network effects. Besides those exogenous factors, the $ERGM_{Degree}$ model includes the network statistics of *network degree (25+)* and $ERGM_{Star}$ includes the network statistics of *network 2-star*. Finally, $ERGM_{Both}$ includes both *network degree (25+)* and *network 2-star* in the model.

Estimates for the ERGMs are shown in Table 5. When comparing the ERGMs including different network effects (i.e., $ERGM_{Degree}$, $ERGM_{Star}$ and $ERGM_{Both}$) to $ERGM_{Null}$, we note that the signs remain the same, whereas the magnitudes of the car attributes vary marginally. The changes are likely due to the correlation introduced by the newly added network structure variables. As for the estimated coefficients of the network structure, the negative sign of the *network degree (25+)*[1] implies that in this particular sedan market, it is less likely for any specific car model to have more than 25 customers. The positive coefficient of *network 2-star* in ERGM indicates a higher probability of graphs with more 2-stars. Since more 2-stars come from nodes of a higher degree, the positive coefficient of this effect implies a higher degree of dispersion, so that there is a subset of car models that are highly popular among customers, compared with others. The running time of different models in R programming with Statnet package are reported based on an Intel i7-11700K CPU with a maximum frequency of 3.6 GHz.

Table 6 shows the estimated results for the *ChoiceSetAll* test case. This corresponds to the choice situation where every customer considers all car models, which is equivalent to knowing the market-level information. In the $ERGM_{Star}$ model, we have included *network 2-star* and *network 3-star* effects. The choice of the network statistics in Test Case 2 is different due to the features of the network under study as well as the convergence issues experienced with the ERGMs (Butts *et al.* 2014). For example, we found that the inclusion of *network degree (25+)* prevented ERGM convergence, and as a result, a different network structure, *network 3-star*, was adopted (see Section 6 for a detailed discussion on the model limitations). It is also observed that the magnitudes of the estimates changed primarily due to the inclusion of network effects. But, the signs of nodal variables are unaffected. This implies that while maintaining the same interpretation of the vehicle features in customer preferences, network effects are important in the formation of the consumer choice network. The predictive performance of these models will be evaluated in Section 4.3. The estimated positive 2-star and negative 3-star coefficients indicate some centralisation through high-degree nodes (i.e., popular cars are more likely to be purchased) but with a cap on the level of that centralisation indicated by the negative 3-star coefficient.

On balance, the sign, significance and interpretation for most car attributes are consistent across the choice set treatments. Broadly, consumers are more likely to choose a car that is more affordable, powerful and with a European brand. In addition, network-based models can model features related to the market structure and thereby provide more information and opportunity for interpretation. Next, we further compare the structures based on predictive performance.

[1] The choice of car degree (25+) is based on the number of total customers in the sample and is also determined by trial and error for model convergence.

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

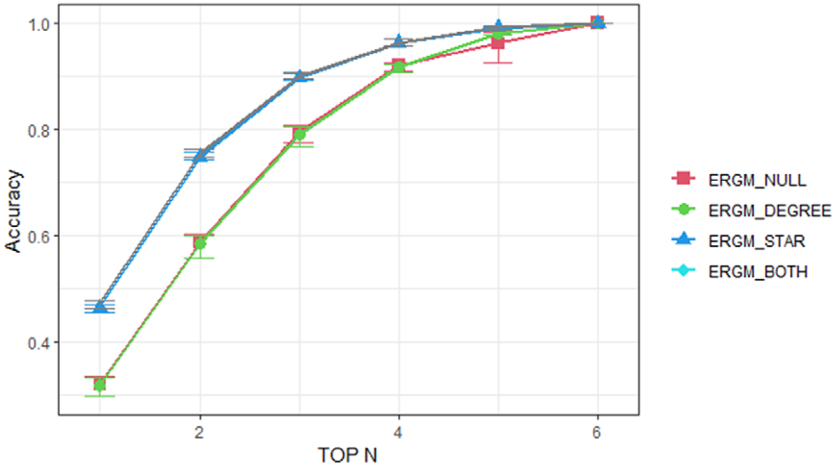


Figure 4. Top-*N* predictions results for three different test datasets (with mean and error bar) by different models in ChoiceSet6 test case.

4.3 Evaluating the ERGM-based choice prediction

The probabilities of customer choices are calculated based on the method introduced in Section 3.1, from which we make the choice prediction. The Top-*N* choice probability and the market share ranking, corresponding to the individual choice prediction and market-level prediction, served as the two evaluation criteria for gauging the predictive power of all the models.

Test case 1: ChoiceSet6

Evaluating the Top-*N* choice probability. Figure 4 shows the average prediction accuracy with error bar of three different test datasets corresponding to the Top-*N* probability. In *ChoiceSet6*, *N* is set from 1 to 6. A higher Top-*N* accuracy indicates a better prediction of the overall trends of customers' preferences over the entire consideration set, not merely for the final choice. For example, the average top 3 accuracy of the $ERGM_{Null}$ model is 0.795, which indicates that the model correctly predicts 3975 customer (i.e., 79.5% of 5000) choices within the model's top 3 predicted choices. We also calculated the area under the Top-*N* curve for each model in Figure 4 in order to assess the overall predictive performance. The baseline $ERGM_{Null}$ model yields an area under curve of 3.92, whereas the $ERGM_{both}$ has the largest area under curve of 4.34, representing a 10.74% improvement over $ERGM_{Null}$.

The inclusion of network structures (i.e., the degree and star effects) in the ERGMs significantly enhances the model prediction, particularly the $ERGM_{Star}$ and $ERGM_{Both}$ models in the top 1 and top 2 predictions. The findings indicate that more detailed network structures could capture latent information in customers' purchase decisions, thereby leading to a better market forecast. As a broader insight, this result implies, in addition to vehicle attributes, that the embedded relations (e.g., competition or popularity) with other car models may also play an essential role in customers' decision-making process. For example, the 'star effect' frequently occurs in the vehicle purchase network, so adding this structure helps explain the network formation process. The $ERGM_{Degree}$ model does not produce

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

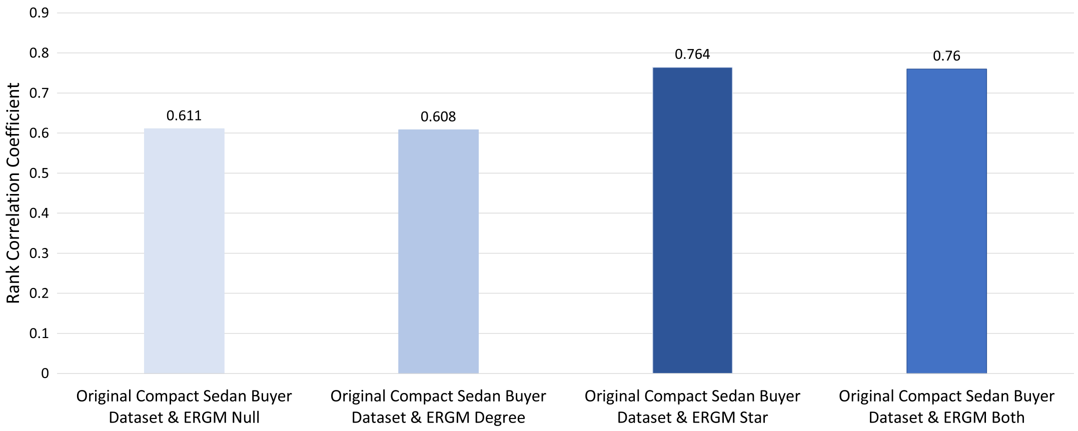


Figure 5. Rank correlation comparison between different models and the baseline value from the original compact sedan buyer dataset.

Table 7. Root mean square errors (RMSEs) between the market share of the original compact sedan buyer dataset and the predicted market share of exponential random graph model base choices

	RMSE
Original compact sedan buyer dataset V.S. $ERGM_{Null}$	0.0104
Original compact sedan buyer dataset V.S. $ERGM_{Degree}$	0.0122
Original compact sedan buyer dataset V.S. $ERGM_{Star}$	0.0118
Original compact sedan buyer dataset V.S. $ERGM_{Both}$	0.0118

significant improvement, indicating that the inclusion of the statistics of *network degree (25+)* is not as effective as the star effect in improving the model's predictive performance.

Evaluating the market share prediction. The predicted market share ranking of each car model is calculated, in which the predicted car model purchases are obtained from the predicted choice probabilities calculated by different ERGMs. Based on the predicted market share, the rank of each car model is obtained, and the resulting rank correlation and RMSEs are simply obtained by comparing it with the market share of the original compact sedan buyer dataset that covers all 84 unique car models. Figure 5 demonstrates the rank correlation coefficients from comparing the original sedan car buying dataset to each of the estimated models, and Table 7 shows the RMSE values.

In Figure 5, the rank correlation coefficient indicates to what extent the predicted popularity of the different models is consistent with that of the original compact sedan buyer dataset. A higher coefficient value means a better prediction of the market share ranking. It is observed that after introducing the 2-star network structure into the ERGM, an improvement of 0.15 in the rank correlation is achieved in both the $ERGM_{Star}$ and $ERGM_{Both}$ models. However, we observe no significant improvement in the $ERGM_{Null}$ and $ERGM_{Degree}$ models. These results

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

are in accordance with the ones observed from the Top-*N* prediction. Specifically, the inclusion of both network structures, that is, the star and degree effects, does not help to further improve the prediction. Instead, the ERGM with only star effect included performs the best among all models. This indicates that accounting for the broader network structures is not universally improving models, and that choosing suitable network structures is critical to constructing well-performing ERGMs. Based on our experience, an in-depth understanding of both the market context and the network structure is needed to help identify a list of potential network structure variables for initial comparison and testing.

For the RMSE evaluation shown in Table 7, the smaller the RMSE value (the closer to zero), the better the prediction. All in all, the models achieve a decent prediction with RMSEs close to 0.01. No superiority of any ERGM is observed. This means that ERGMs, with the inclusion of specified network structures, perform better than $ERGM_{Null}$ in predicting the overall market share ranking but achieves comparable, or slightly worse prediction for the market share percentages.

Test case 2: ChoiceSetAll

In the *ChoiceSetAll* treatment,[2] every individual customer will have the same choice set that consists of the same number of alternatives, that is, 281 vehicles, from the entire (simulated) market. Figure 6 shows that $ERGM_{Null}$ yields a better overall prediction accuracy with an area under curve of 214.30, whereas $ERGM_{Star}$ shows

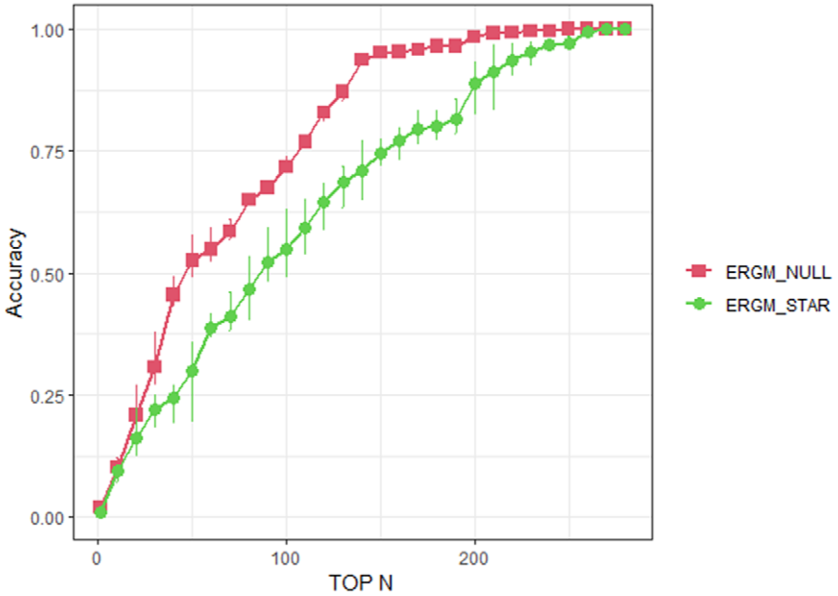


Figure 6. Top-*N* predictions results for three different test datasets (with mean and error bar) by different models in ChoiceSetAll test case.

[2] Since no network structure characteristics are included in the $ERGM_{Null}$ model and every customer has the same universal choice set, the prediction yields identical choice prediction results. Therefore, the criterion of market share is not applicable here.

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

a lower accuracy measured by area under curve of 180.37. This implies that when modelling the scenario in which each customer faces all the car alternatives, adding the star-network effect does not help to improve the model performance. It could be the reason that in such a *ChoiceSetAll* scenario, each customer has already had access to the market-level information from the choice set. So, adding additional network structures does not further improve the predictive power of the model. This calls for caution when including network structures in the model for prediction purposes, as the types of network structures, the potential correlations between network effects and nodal attributes, the assumptions on choice scenarios, the size of choice set and the size of the test data could all influence the model performance. An evaluation of the model's prediction accuracy benchmarked on the null model is always recommended before the final adoption of a model with network structures included.

5. Implications for engineering design

In this section, we present an improved DBD framework that integrates the ERGM-based choice model. The utility of the ERGM-based DBD framework is demonstrated in Section 5.2 by comparing the results of the MNL choice model (a traditional DCM commonly used in the DCA literature) with the ERGM-based choice model.

5.1. ERGM-based choice analysis for design

Based on the classic DBD framework (Chen *et al.* 2012), an improved DBD framework that integrates the ERGM-based choice model is proposed in Figure 7. In this framework, the grey boxes indicate the entities and events that belong to the original DBD framework. For example, the expected utility, in the ERGM-based framework, is still the key entity to merge the marketing and engineering domains into a single enterprise-driven decision-making framework.

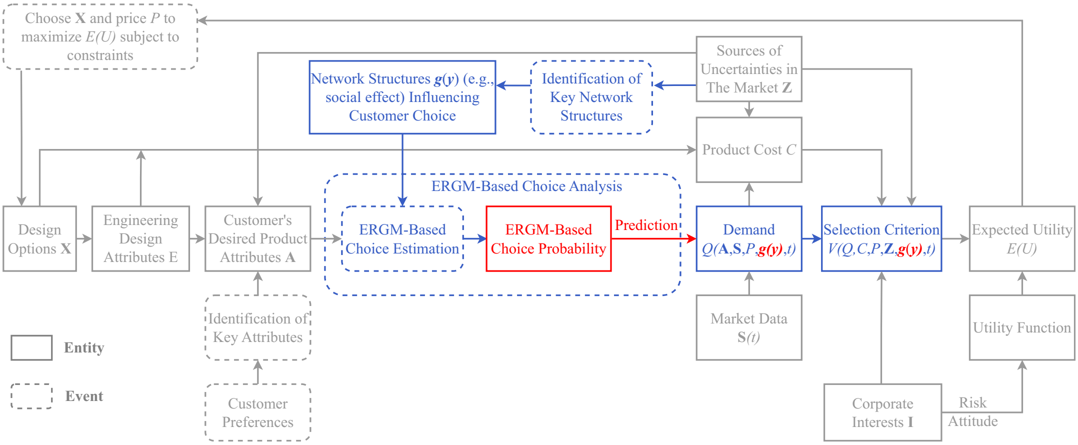

Choose X and price P to maximize E(U) subject to constraints
Sources of Uncertainties in The Market Z
Network Structures g(y) (e.g., social effect) Influencing Customer Choice
Identification of Key Network Structures
Product Cost C
ERGM-Based Choice Analysis
Design Options X
Engineering Design Attributes E
Customer's Desired Product Attributes A
ERGM-Based Choice Estimation
ERGM-Based Choice Probability
Prediction
Demand Q(A,S,P,g(y),t)
Selection Criterion V(Q,C,P,Z,g(y),t)
Expected Utility E(U)
Identification of Key Attributes
Market Data S(t)
Utility Function
Entity
Event
Customer Preferences
Corporate Interests I
Risk Attitude

Figure 7. Decision-based design framework enhanced by network models. The grey boxes are the entities and events from the referred decision-based design framework; the coloured boxes belong to the proposed ERGM-based choice analysis.

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

The coloured boxes indicate the place where the traditional discrete choice analysis is replaced by the ERGM-based choice analysis. As shown in the figure, the ERGM-based choice analysis considers not only the effect of product attributes on customer choices but can also the effect of network statistics $\boldsymbol{g(y)}$. These network statistics can be used to model the influence of social effects, product relations and any market effects that capture the interactions between customers and products on customers' choice behaviour. With the inclusion of these network features, it is expected that a higher prediction accuracy of the choices and, therefore, the demand (Q) can be achieved. This hypothesis will be validated in the next section via a comparative study. The design implication of choice modelling is embodied through the successful prediction of market demand that can be used to construct the expected utility function ($E(U)$). From there, an optimisation problem can be formulated to find the optimal design options X and the price P that maximise $E(U)$ subject to constraints, such as cost. Since the formulation of such a design problem has been extensively studied (Wassenaar & Chen 2003; Wang, Kannan & Azarm 2011; Shin & Ferguson 2017; Yip, Michalek & Whitefoot 2019), we do not demonstrate it again in this paper. Instead, we will focus on validating the power of the ERGM-based choice model in predicting choices and demand, using the classical MNL choice model as the benchmark. It should be noted that, because of the introduction of network effects in this new DBD framework, it is now possible to test how different social relations, product associations and market effects would influence the design of product attribute.

5.2 Comparing the multinomial logit DCM and ERGM-based choice model

In this comparative study, we choose the commonly used MNL choice model as a representative DCM and use it as the baseline to compare against the model $ERGM_{Both}$. We compare their model estimates as well as their prediction accuracy of choices.

Table 8. Estimated results of ChoiceSet6 test case with the comparison between $ERGM_{Both}$ and DCM

Explanatory variables	DCM_{MNL}	$ERGM_{Both}$
Price	−5.113551***	−3.02861***
Fuel consumption	0.416578***	−0.03438
Power	2.230066***	1.59093***
Brand origin (the US)	3.893254***	1.22953***
Brand origin (Europe)	5.822052***	2.20565***
Brand origin (Japan)	4.184239***	1.21286***
Brand origin (Korea)	3.542386***	1.16191***
Network degree (25+)		1.20514**
Network 2-star		0.00583***

Note: The code *** indicates the 0.001 level of significance and ** indicates the 0.005 level of significance.

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

Comparing the model estimates

The estimated parameters of the DCM and $ERGM_{Both}$ in *ChoiceSet6* test case are shown in Table 8. For a fair comparison, we specify the same set of exogenous variables (i.e., car attributes) for both DCM and ERGM. The first observation is that the magnitudes of the estimates differ notably between the DCM and $ERGM_{Both}$. But, the signs of the estimates of these two models are mostly similar, except for fuel consumption. The sign coefficient for fuel consumption is expected to be negative, indicating that a car with lower fuel consumption is more desirable. The results show that the estimate of fuel consumption in $ERGM_{Both}$ is more interpretable. The counter-intuitive coefficient of fuel consumption in the MNL model could be caused by the collinearity between input model terms (e.g., high fuel-consumption cars usually also have higher power). Also, we noticed that in the MNL model, car models with brands from Japan are more favourable than those from the United States, which is the opposite case in ERGM, where car models with brands from the United States are slightly more preferred. We argue that the interpretation of the ERGM model is more reasonable since it takes into account the overall network structure (i.e., market-level information).

Comparing the predictive performance

Still taking *ChoiceSet6* as an example, as shown in Figure 8, it is observed that the Top-*N* prediction accuracy of $ERGM_{both}$ is higher than that of DCM, indicating that ERGM outperforms MNL when predicting the overall trends of customers' preferences over the entire consideration set. Regarding the prediction of market share ranking, the rank correlation coefficient for $ERGM_{both}$ is 0.76, which is higher than DCM (0.607), indicating the superiority of $ERGM_{both}$ in predicting the popularity of products in the overall market. In terms of the prediction of market share percentages, both $ERGM_{both}$ (0.0118) and DCM (0.0104) achieve about equal prediction with RMSEs close to 0.01, but no superiority of $ERGM_{both}$ is observed.

In short, the results reveal that, in addition to having better interpretability, ERGM with network structures ($ERGM_{both}$) also outperforms DCM in predictive powers of both the individual choice behaviours and the overall market shares, measured by the Top-*N* choice prediction accuracy, rank correlation of the market

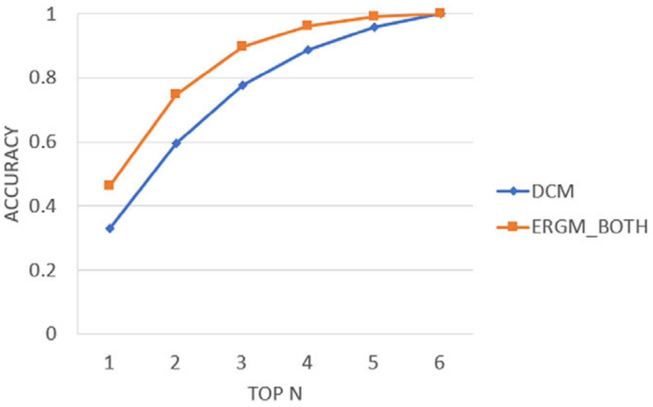


Figure 8. Top-*N* predictions for DCM and $ERGM_{Both}$ model in ChoiceSet6 test case.

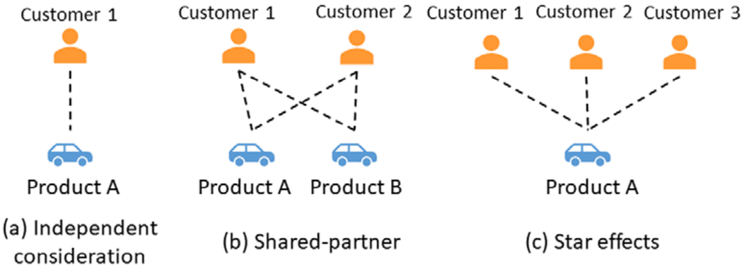
https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

share as well as the market percentage RMSE. This outperformance further validates the rationality of integrating ERGM analysis into DBD framework as shown in Figure 7.

5.3 Discussion on theoretical foundation of DCM and ERGM-based choice model

As stated in Section 1, the basic forms of DCMs have restrictions due to their assumptions on the independencies between alternatives and the rationality of customers, and the network-based approach addresses these limitations by treating the choice prediction as a link prediction problem so that specific network models, that is, ERGM in this study, can help relax those assumptions. This is because in contrast to the usual general linear model, which assumes that customers' considerations are independent (Figure 9a), network models assume that the formation of links is dependent on other links in the network. For example, the link formed between Customer 1 and Product A may be because Customers 1 and 2 both considered Product B, and meanwhile Customer 2 also considered Product A (the case shown in Figure 9b), or may be because both Customers 2 and 3 considered Product A (the case shown in Figure 9c). In ERGM, different network configurations, such as the 'shared-partner' structures (e.g., cars being considered by common customers), as shown in Figure 9b, and 'star effects' (e.g., the cars being considered by many customers), as shown in Figure 9c, can model those interdependent effects.

The differences between the two models' estimates and outperformed prediction ability of ERGM are primarily caused by the distinction between the calculation of choice probabilities in DCM and ERGM and their resulting maximum likelihood estimation processes. In DCM, the choice probability is calculated within the given choice set. However, in ERGM, the calculation takes into account the entire network structure because of the inclusion of network statistics. This sheds light on the interpretation of the decision-making assumptions underlying the two models. In DCM, the model estimation assumes that each customer evaluates their own choice set and makes a final choice by picking the alternative with the maximum utility. That is, any comparison of utilities is done within the choice set, and customers do not refer to the alternatives outside that set. However, in ERGM, the linking probability between two nodes is calculated based on the information from the entire network, that is, the market. So, ERGM assumes that customers are aware of market-level information and while they are making a



(a) Independent consideration

(b) Shared-partner

(c) Star effects

Figure 9. The interdependency assumption underlying the network-based models.

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

choice decision, such information will consciously or unconsciously affect their choice behaviours. This is what Shocker *et al.* (1991) mentioned the *awareness set*, which can provide additional cue information in memory to customers when they make decisions. Arguably, the configurations in network models represent a more realistic decision process where consumers are likely to be aware of all the car models in the entire set, and ERGMs can provide an avenue to examine that awareness effect on choice while also improving their own predictive abilities.

6. Conclusion

In this paper, we develop a novel ERGM-based choice prediction approach with the ability to compute individual choice probabilities in a given choice set. This approach, on the one hand, explore the customer choice prediction power of ERGM in a bipartite network setting, that is, the incorporation of network structures to improve the choice prediction accuracy. On the other hand, this approach is a key bridge between ERGM-based choice analysis and the concept of 'utility' in traditional DCMs when extending the classic DBD framework (Chen *et al.* 2012) to the DBD framework enhanced by network models. The utility of this approach is validated by using a vehicle market system as a case study. According to the results, we observe that the inclusion of network structures in ERGM yields a significant improvement in predicting individual choices as quantified by the Top-*N* choice prediction as well as the market share rank correlation validation. An important caveat to note for each model structure is that the predictive performance of ERGM is related to the selection of specific network structures. Therefore, an in-depth understanding of the market structure (i.e., network formation) will be key to identifying the set of most suitable and effective network structures for ERGM modelling. Relatedly, this choice will likely vary with the decision context and sample analysed.

Based on this study, our future work unfolds in two directions. First, a multidimensional choice network is under investigation, where more complex relations at each layer of nodes, for example, product associations at the product node layer and social interactions at the customer layer, are allowed, so that it provides a framework for researchers to study how the interdependencies of multiple layers of nodes can influence customers' choice behaviours, and thus helps to forecast the market share. Second, the authors see the need of developing a feature engineering approach for automatic network structure identification and selection in support of constructing ERGMs with network structural effects.

Financial support

The authors acknowledge the financial support from NSF-CMMI-2005661 and 2203080, the Ford-Northwestern Alliance Project and the Vice Chancellor for Research and Innovation Seed Grant from the University of Arkansas. Partial financial support was provided to Stathopoulos from NSF-CAREER-1847537.

References

Ahmed, F., **Cui, Y.**, **Fu, Y.** & **Chen, W.** 2021 A graph neural network approach for product relationship prediction. In *'47th Design Automation Conference (DAC)', Proceedings of*

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

the ASME Design Engineering Technical Conference. American Society of Mechanical Engineers.

Bi, Y., **Qiu, Y.**, **Sha, Z.**, **Wang, M.**, **Fu, Y.**, **Contractor, N.** & **Chen, 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.

Bi, Y., **Xie, J.**, **Sha, Z.**, **Wang, M.**, **Fu, Y.** & **Chen, W.** 2018 Modeling spatiotemporal heterogeneity of customer preferences in engineering design. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 51753)*, p. V02AT03A050. American Society of Mechanical Engineers.

Butts, C. T., **Morris, M.**, **Krivitsky, P. N.**, **Almquist, Z.**, **Handcock, M. S.**, **Hunter, D. R.**, **Goodreau, S. M.** & **de Moll, S. B.** 2014 *Introduction to exponential-family random graph (ERG or p*) modeling with ergm. European University Institute, Florence.* http://cran.r-project.org/web/packages/ergm/vignettes/ergm.pdf.

Chai, T. & **Draxler, R. R.** 2014 Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development* **7** (3), 1247–1250.

Chang, D. & **Chen, C.-H.** 2014 Understanding the influence of customers on product innovation. *International Journal of Agile Systems and Management* **7** (3–4), 348–364.

Chen, W., **Ahmed, F.**, **Cui, Y.**, **Sha, Z.** & **Contractor, N.** 2020 Data-driven preference modelling in engineering systems design. In *Handbook of Engineering Systems Design* **(ed.** A. Maier, J. Oehmen & P. E. Vermaas), pp. 1–34. Springer International Publishing; doi:10.1007/978-3-030-46054-9_15-1.

Chen, W., **Heydari, B.**, **Maier, A. M.** & **Panchal, J. H.** 2018 Network-based modeling and analysis in design. *Design Science* **4**, e16.

Chen, W., **Hoyle, C.** & **Wassenaar, H. J.** 2012 *Decision-Based Design: Integrating Consumer Preferences into Engineering Design*. Springer Science & Business Media.

Chen, C.-H., **Khoo, L. P.** & **Chen, N.-F.** 2015 Consumer goods. In *Concurrent Engineering in the 21st Century*, pp. 701–733. Springer.

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

Cui, Y., **Ahmed, F.**, **Sha, Z.**, **Wang, L.**, **Fu, Y.**, **Contractor, N.** & **Chen, W.** 2022 A weighted statistical network modeling approach to product competition analysis. *Complexity* **2022**, 9417869.

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. S., **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. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 58127)*, p. V02AT03A039. American Society of Mechanical Engineers.

Handcock, M. & **Hunter, D.** 2018 Reference manual of package ergm. https://cran.r-project.org/web/packages/ergm/ergm.pdf.

Harris, J. K. 2013 *An Introduction to Exponential Random Graph Modeling*, Quantitative Applications in the Social Sciences (Vol. 173). Sage Publications.

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

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

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.

Hoyle, C. J. & **Chen, W.** 2009 Product attribute function deployment (PAFD) for decision-based conceptual design. *IEEE Transactions on Engineering Management* **56** (2), 271–284.

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), 121007.

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), nihpa54860.

Kang, N., **Ren, Y.**, **Feinberg, F. M.** & **Papalambros, P. Y.** 2016 Public investment and electric vehicle design: a model-based market analysis framework with application to a USA–China comparison study. *Design Science* **2**, e6.

Luce, R. D. 2012 *Individual Choice Behavior: A Theoretical Analysis*. Courier Corporation.

MacDonald, E. F., **Gonzalez, R.** & **Papalambros, P. Y.** 2009 Preference inconsistency in multidisciplinary design decision making. *Journal of Mechanical Design* **131** (3), 031009.

McFadden, D. & **Zarembka, P.** 1974 Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, pp. 105–142. Academic Press.

McFadden, D. 1973 *Conditional logit analysis of qualitative choice behavior*. In *Frontiers in Econometrics*, pp. 105–142. Academic Press.

Morris, M., **Handcock, M. S.** & **Hunter, D. R.** 2008 Specification of exponential-family random graph models: terms and computational aspects. *Journal of Statistical Software* **24** (4), 1548.

Mottini, A., **Lhéritier, A.**, **Acuna-Agost, R.** & **Zuluaga, M. A.** 2018 Understanding customer choices to improve recommendations in the air travel industry. In *RecTour@ RecSys*, Association for Computing Machinery pp. 28–32.

Schober, P., **Boer, C.** & **Schwarte, L. A.** 2018 Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia* **126** (5), 1763–1768.

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. In *Proceedings of the Design Society: International Conference on Engineering Design (Vol. 1)*, pp. 3831–3840. Cambridge University Press.

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**, 2753638.

Sha, Z., **Saeger, V.**, **Wang, M.**, **Fu, Y.** & **Chen, W.** 2017a 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.** 2017b Modeling product co-consideration relations: a comparative study of two network models. In *DS 87-6 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 6: Design Information and Knowledge, Vancouver, Canada, 21–25 August 2017*, pp. 317–326. The Design Society.

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), 021402 https://www.designsociety.org/publication/39495/DS+87-6+Proceedings+of+the+21st+International+Conference

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

+on+Engineering+Design+(ICED+17)+Vol+6:+Design+Information+and+Knowledge,+Vancouver,+Canada,+21-25.08.2017.

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.

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

Train, K. 1986 *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand (Vol. 10).* MIT Press.

Wang, M., **Chen, W.**, **Huang, Y.**, **Contractor, N. S.** & **Fu, Y.** 2015 A multidimensional network approach for modeling customer–product relations in engineering design. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 57175).* American Society of Mechanical Engineers.

Wang, M., **Chen, W.**, **Huang, Y.**, **Contractor, N. S.** & **Fu, Y.** 2016a Modeling customer preferences using multidimensional network analysis in engineering design. *Design Science* **2**, e11.

Wang, Z., **Kannan, P.** & **Azarm, S.** 2011 Customer driven optimal design for convergence products. *Journal of Mechanical Design* **133** (10), pp. 379–394 https://doi.org/10.1115/1.4004977.

Wang, M., **Sha, Z.**, **Huang, Y.**, **Contractor, N.**, **Fu, Y.** & **Chen, W.** 2016b Forecasting technological impacts on customers' co-consideration behaviors: a data-driven network analysis approach. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 50107)*, p. V02AT03A040. American Society of Mechanical Engineers.

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**, e9.

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

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), 031112.

Yip, A. H., **Michalek, J. J.** & **Whitefoot, K. S.** 2019 Implications of competitor representation on optimal design. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 59186)*, p. V02AT03A045. American Society of Mechanical Engineers.

Appendix. Validating the representativeness of the training datasets

To check if the market share of the sampled training dataset (i.e., the dataset consists of 5000 car buyers) effectively represents the market share of the compact sedan in the original compact sedan buyer dataset (i.e., the dataset including all 18,054 car buyers), the market shares from both datasets are compared. After obtaining and ranking the market share of every compact sedan from the original dataset, the top 10 most popular compact sedan car models are identified (see the blue bars in Figure A1). In the same figure, the orange bars represent the market share and the ranks of those top 10 car models obtained from the training dataset.

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press

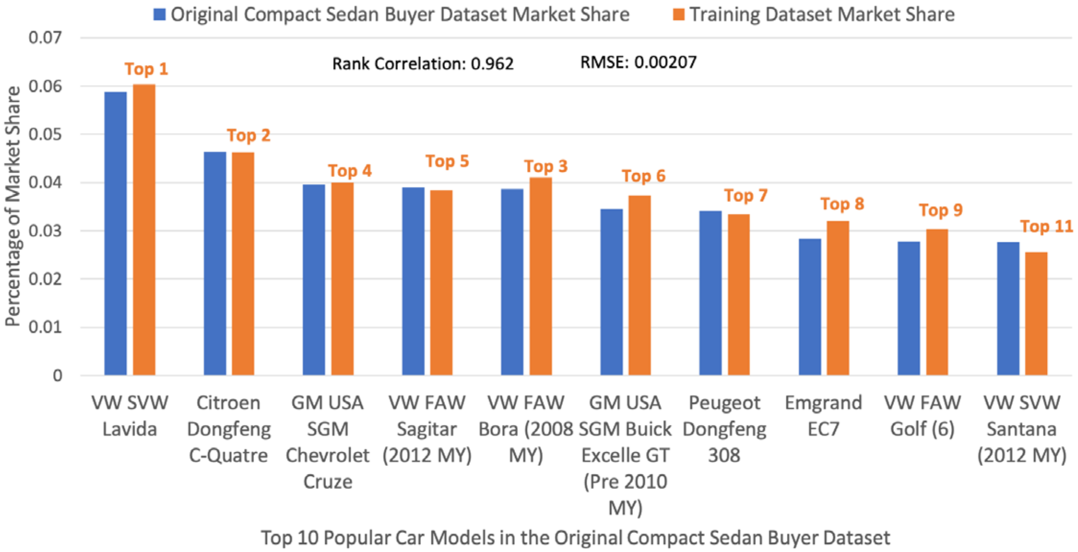


Figure A1. Market share comparison of the original compact sedan purchased dataset and the training dataset.

As shown in Figure A1, the rank of the market share of the top 10 compact sedans from the training dataset has a high rank correlation (0.96) with the original data, indicating that the training dataset is a good representation of the original compact sedan buyer dataset. This conclusion is confirmed by the low RMSE value of 0.002. To further evaluate the market share correlation and variation between the training dataset and the original compact sedan buyer dataset, the rank correlation and market share RMSE of the total set of 84 unique compact sedan car models are calculated. The results are 0.94 and 0.0027, respectively. Therefore, the representativeness of the training datasets is confirmed.

https://doi.org/10.1017/dsj.2023.4 Published online by Cambridge University Press