Customized Individual Promotions: Model, Optimization, and Prediction

For a retailer, running personalized promotions is a means to overcome the negative effects of promotions offered to all customers (e.g., stockpiling effects on the customer side). These customized promotions rely on the heterogeneous preferences of individuals for products within a category, and their different sensitivities to price discounts. We consider the problem of predicting individual customer responses to promotion decisions over a category of products, and based on the predictions, optimizing the portfolio of products to put on promotion for a particular individual.

The data sources required by our methodology include historical purchase transactions data tagged by customer ID, information about the assortment available for the product category of interest at the moment of purchase, and the identification of the products that were on promotion by that time. In our model each customer is represented by a partial order or directed acyclic graph (DAG). The DAG is constructed from a set of rules that account for the revealed preferences of each customer through the history of past purchases. In a DAG, each node represents a product, and each arc represents a pairwise preference relation. While processing the data source in the DAG construction phase, a customer may exhibit an apparent inconsistency in her purchase behavior that may imply the creation of a cycle. Then, we run a de-cycling procedure based in a mixed-integer programming (MIP) formulation in order to keep a maximal number of arcs to describe consistent customer preferences.

On the theoretical side, we provide tractable bounds to compute both the likelihood of partial orders (which in general is a #P-hard problem) and the purchase probabilities from these DAGs.

Next, taking the collection of DAGs representing the customer basis as input, we calibrate a single class MNL model and a latent class MNL model. On the side, we also estimate two state-of-the-art parametric models: latent-class multinomial logit (LC-MNL), and random parameters logit (RPL) accounting for panel data, but without using the aforementioned DAG-based underlying structure. Finally, our DAG-based MNL variants and both benchmarks were tested for designing personalized promotions. Via a MIP formulation, we consider the problem where the offer set faced by the customers in each store visit is given and the decision variables are which products to promote in order to maximize expected revenues. Numerical experiments on real-world panel data for a grocery chain across 27 product categories. We observe that our approach allows more accurate, fine-grained predictions for individual purchase behavior compared to the state-of-the-art benchmarks. For the performance score functions that we computed, our DAG-based models could outperform the benchmarks by more than 10%. Furthermore, our MIP for running personalized promotions can increase revenues by 26% (on average) over the current practice reflected in the dataset.

**Key words**: choice models, promotions optimization, integer programming, assortment planning

**History**: This paper was first submitted.
1. Introduction

The real time access to customized data and new technology allows retail stores to make personalized operational decisions. In particular, technological development opens new opportunities for the retailers to run customized promotions in the offline setting, where 85% of sales take place (Pounder 2015). For example, Lowes Foods, the grocery store chain with around 100 stores in North Carolina, South Carolina and Virginia, launched personalized promotions in May 2016 (Denman 2016). It became possible due to the integration of Lowes Foods’ loyalty program with Circular 2.0 platform created by Unata. As a result, every registered store guest receives the most relevant deals, based on their purchasing history, through the mobile phone. In addition, the platform allows to measure the conversion rate from online views to in store purchases. According to Mike Minasi, president of marketing in a large retail chain Safeways, shelf prices will become significantly less relevant over several years for a large part of the shoppers (Kharif 2013). In 2013, about 45% of Safeway’s sales came from specialized offers that customers could get through PC or mobile applications.

First of all, customized promotions can be run through personal shop assistant (PSA) device which is programed to identify a customer by her loyalty card inserted into the PSA. In addition, radio-frequency identification (RFID) tags in the products can help the retailer to get additional information about the customer, for example, to detect the moment the consumer takes the item from the shelf, and to use this information to improve customized promotions. Another alternative to personalize consumers is to use electronic price tags that change the price based on the customer identification. This technology has already been used by B&Q retail chain (Pounder 2015). A number of retail stores like Macy’s, Marsh supermarkets, Gamestop, and mall developers such as Simon Property Group and Macerich have tested a beacon-based technology that allows to send relevant messages to the targeted customers (Korber 2015). Beacon technology has a special feature to launch the application downloaded by the customer, even if it is not open in the phone. Simon Property Group installed about 4800 beacons over 192 malls, where customers can be targeted by Simon app. This technology both improves sales and induces the customers’ desire to make a purchase. For the top 100 retailers in 2015, beacon-based sales were around $4 billion. According to the research done by beacon platform Swirl, about 73% of customers, that made purchases triggered by beacon, responded that this technology increased the likelihood of buying an item during the store visit, and 60% of respondents said that this targeted marketing makes them visit the store more often.

Even though past attempts to customize marketing did not fully use the potential of the cutting edge technology discussed above, some companies always tried to be creative to address the need
of their customers and to strategically market promotions to their audience. For many years, a few big retailers mailed customized packets of paper coupons to the customers as a part of the loyalty program once a month or once a quarter. Nowadays similar discounts are offered online through a browser or mobile application at least once a week. For example, in 2013 customers of Kroger started to receive up to 150 coupons a week through the mobile app. Metro’s chief marketing officer claims that the more frequently retailers communicate with the shoppers, the more sales they generate. In September 2013, Metro started its digital Metro&Moi portal and a mobile application that provides personalized coupons on a weekly basis.

Personalized promotions can be used by retailers as a perfect tool for price discrimination between different customers. Another benefit of customized promotions for the retailer is the reduction of competition (Kharif 2013). Because targeted promotions are directed to the consumer, it is hard for the competitors to track it. Not only retailers benefit a lot from individualized promotions, but also consumers themselves appreciate services accompanied with personalization more than they dislike sharing the personal information about their purchasing habits (Farhnam 2013). Personalization induces stronger customer relationship and drives sales. According to Accenture, more than 60% of customers want to participate in customized promotions and explore real time deals. Thus, retailers that do not use personalization with customers lose additional opportunities to build a stronger relationship with consumers and make them more loyal to the retail store.

Given the importance of customized promotions in the grocery industry, it is not surprising that a few supermarkets utilize services provided by consumer science companies. For example, Catalina Marketing provides customized coupons to retailers and brands by tracking the purchase activity of more than 230 million of shoppers in US. Dunnhumby is another customer data science company that tracks more than billion items sold by retailers each week. Their software detects customer’s buying pattern, favorite brands, and purchased items to create customized offers for the products that customers bought or are likely to buy. This software provides personalized offers to many retailers including Metro and Kroger. Receiving offers through a website or mobile app, customers redeem them via swiping the loyalty cards at checkout. Simon Hay, chief executive officer of Dunhumby, suggests that the number of individualized offers will keep increasing from current 5% of total offers to more than 30% in several years, where the main driver will be an increased use of mobile devices and their applications.

Our main contribution is to propose a back-to-back technique to run personalized promotions from the retailer’s perspective. Our algorithm finds the set of products to put on promotion for every individual given their purchasing history and the set of available items at the store. Each promotion has its own pros and cons and serves a particular objective of a retailer. There might be several main effects of promotions: switching effect (e.g., store switching, category switching, or brand
switching), stockpiling effect, consumption stimulation effect, new customers attraction effect, flash sale effect (i.e., quick disposal of unwanted items), loyalty promotion to keep existing customers (e.g., store loyalty, category loyalty or brand loyalty). Running personalized promotions is a very hard modeling and empirical problem even with enough historical transaction data provided by the retail stores. In the present paper, we focus on the brand switching effect of personalized promotions and propose a systematic way to induce a brand switching effect to increase retailer’s revenue. To this end, personalization plays an important role, since there is no reason to promote an item to customers who already buy it on a regular basis. At the same time, offering the same deal for all the consumers, the retailer is likely to trigger the stockpiling effect from brand loyal customers, which could negatively affect the promotion effectiveness. As a result, it is more efficient for the retailer to target each consumer individually based on their purchasing history and sensitivity to promotions.

1.1. Summary of The Results and Contribution

In this section we summarize contributions of the paper and provide an overview of the main results:

- We propose and analyze a nonparametric choice-based demand model that explicitly accounts for promotions. We extend the nonparametric partial-order-based choice model (Jagabathula and Vulcano 2017) to capture the promotion effect. The proposed model belongs to the family of Random Utility Maximization (RUM)-based choice models. In particular, we consider a fixed set of customers making repeated purchases from a specific category of substitutable products over a finite time horizon. Every customer is described by a partial preference which can be represented by a directed acyclic graph (DAG), where there are two nodes per product (one representing the product at its full price, and one representing its promoted version) and arcs representing the pairwise preferences among the products. In each store visit, every customer samples a full preference list in accordance with her partial order according to a pre-specified distribution, and chooses the highest rank item within her consideration set. From sales data, we only observe the collection of revealed preferences for every individual: the set of available products at the time of purchase and the chosen product. Consequently, based on revealed preferences of a customer through her purchasing patterns, we dynamically build customer’s DAG. Note that customer’s consideration set is unobserved, which makes the estimation problem even more challenging.

- We introduce a data-driven approach to account for customers’ consideration sets. The challenge is that ignoring consideration sets results in adding spurious edges (i.e., spurious comparisons) to the DAG in building a customer’s partial order. Specifically, incorrectly assuming a large consideration set (e.g., consideration set consists of all the available items in the store) increases both
the number of correct and spurious edges in the DAG. On the other hand, incorrectly assuming sparsity in consideration set formation decreases the number of both types of edges. Thus, it is worth finding a trade-off between maximizing the number of correct pairwise comparisons (e.g., in order to provide a more accurate estimation of the underlying model) and minimizing the number of spurious pairwise comparisons (e.g., in order to obtain a less biased estimate of the underlying preference distribution). We address this challenge by accounting for considerations sets in a data driven way while building a customer’s partial order.

- We quantify the predictive accuracy gains in the choice process from modeling promotion effects and accounting for consideration sets in a data-driven fashion. We fit a MNL model to estimate probabilistic distributions over preference lists consistent with the partial orders. In our empirical analyses, we focus on real-world panel data on the sales transactions of 27 grocery categories across two big U.S. markets in 2007. The extensive empirical studies demonstrate that our approach to account for promotion effects results in more precise, fine-grained predictions for customer choice behavior in comparison with state-of-the-art benchmarks that also incorporate promotion effects. In particular, we obtain up to 14% improvement in prediction accuracy on average across 27 product categories.

- We derive bounds on the estimation and prediction guarantees for the partial-order-based choice model. Obtaining the estimate of the marginal distribution for partial preferences under the MNL model requires computing the likelihood of a DAG which is known to be a #P-hard problem. Therefore, we use an approximate solution to efficiently compute the likelihood of the DAG and the probability for the customer to choose a specific product from an offer set conditioned on her partial order. In the present paper, we derive lower and upper bounds relative to the exact likelihood of partial orders and to the exact distribution for partial preferences under the MNL model.

- We propose a methodology for optimizing personalized promotions, and test it on real-world panel data. In the spirit of the operations-related literature, the defined nonparametric choice-based demand estimation model is used as an input for personalized promotion optimization framework to improve the retailers revenue. Our approach to run personalized promotions allows intuitive and illustrative interpretation of the resulting customized promotions policy which is a very appealing property for the retail industry, where a lot of supermarkets still employ a manual process based on rule of thumb and past experience in order to decide price promotions (Cohen et al. 2017a). We observe that our methodology to optimize personalized promotions improves the revenue of the retailers by more than 26% for some product categories based on the real world panel sales transactions data.
1.2. Related Literature

In this paper, we contribute to the literature in both operations and marketing. First of all, our work is related to the stream of research that focuses on choice-based demand estimation for a set of substitutable products using sales transactions and product availability data. Choice-based demand estimation models, which is a current trend in both the academia and the industry practice, significantly outperform traditional independent demand models in explaining customers’ purchasing patterns especially when product variability varies over time as an outcome of either inventory stockouts or deliberate scarcity by retailer (Ratliff et al. 2008, Newman et al. 2014, Vulcano et al. 2010, Dai et al. 2014).

The use of discrete choice models to estimate consumer preferences dates back to the seminal paper by Guadagni and Little (1983), where they calibrated parsimonious multinomial logit model (MNL) on 32 weeks of purchases of regular ground coffee by 100 households. The MNL model is a widely used example of parametric choice models, which assumes a specific closed form relationship between available alternatives with their attributes, and purchase probabilities, through finite-dimensional parameter space. Even though the MNL model has received a great deal of attention among scholars and practitioners because of its tractability and reasonably well predictive performance, it exhibits the independence from irrelevant alternatives (IIA) property, which limits its practical applicability. In models with IIA property the impact of one alternative on another one does not depend on which other alternatives are present in the choice set. More complex parametric models such as the nested logit (NL) or random parameter logit (RPL) overcome this IIA restriction by accounting for customer’s heterogeneity in modeling substitution patterns (Ben-Akiva and Lerman 1985, Train 2009). These models are widely used in marketing when there is access to scanner panel data. A more detailed discussion of choice modeling in the marketing field can be found in the surveys by Chandukala et al. (2008) and Wierenga et al. (2008). More sophisticated parametric models with many nests and latent classes may be able to capture customer’s heterogeneity and result into better predictive performance by reducing the specification error. However, as a result of estimation error those models can be significantly less precise than parsimonious single class MNL model when the modeler faces a data scarcity problem. Thus, there is a trade-off between specification and estimation errors in choosing the choice model specification from parametric class. A distinguishing feature of parametric models is the ability to incorporate covariates, such as product attributes and price, which can further boost the predictive power of the model. Moreover, attraction-based parametric models (e.g., MNL, latent class MNL, random parameter logit (RPL)) are convenient to extrapolate choice predictions to new alternatives that have not been part of historical sales transactions data and to predict the impact of changes
in product attributes (e.g., price) on choice frequencies. However, any parametric model requires assumptions about the structure of customers’ preferences and the relevant covariates that influence it, which results in an ad hoc way of calibration based on expert judgment and trial-and-error testing to determine an appropriate specification.

In the present paper we consider a nonparametric approach to model choice behavior of customers, which complements the former class of models discussed above and extends it to a wider range. The nonparametric choice models specify customer types defined by their rank orderings of all alternatives in the product category, and every customer is assumed to purchase the available product with the highest ranking in her preference list, or to leave without making any purchase. Providing a greater flexibility in modeling customer choices, this class of models might be too permissive and suffer from nonidentifiability (Sher et al. 2011), which is their main drawback. Nevertheless, nonparametric approaches to choice modeling, where the choice probability arises from a sparse distribution over preference lists (Rusmevichientong et al. 2006, Farias et al. 2013, van Ryzin and Vulcano 2014), have been gaining a lot of attention within operations as the result of the increased availability of sales transaction data. To this end, Farias et al. (2013) present a robust estimation methodology to predict expected revenues as a function of offer sets using sales transaction data. van Ryzin and Vulcano (2014) develop a “market discovery” algorithm, which starts with a parsimonious set of customer types and iteratively generates new types that increase the data log-likelihood value. The paper by Jagabathula and Vulcano (2017) is especially relevant to our work. They consider the general nonparametric choice model where the customers are represented with partial orders of preferences. Motivated by marketing literature on consideration sets definitions (Bettman et al. 1998, Hauser et al. 2009), the authors assume that individuals form their consideration sets before making a purchase. In each store visit, a customer samples a full preference list of items in the product universe (along with the no purchase alternative) in accordance with her partial order, forms a consideration set, and then buys the product among the considered ones with the highest rank in the preference list. Their approach accounts for consideration sets of customers by explicitly modeling behavioral biases that are common in customer choice behavior (e.g., inertial model). Our contribution to their study is three-fold. Methodologically, we extend the proposed nonparametric choice model so that it explicitly includes promotions, and based on this model, we present a framework to optimize personalized promotions. In addition, we propose an alternative way to account for customer’s consideration sets. Our approach is data driven and provides a greater flexibility to model customers’ consideration sets without imposing any prior beliefs on individuals’ consideration sets formation. From a theoretical perspective, we develop tractable analytical lower and upper bounds relative to the exact likelihood of the partial orders under MNL model, since it is known that computing the likelihood of the general
partial order is a \#P-hard problem. In addition, we derive tractable analytical lower and upper bounds relative to the exact marginal distribution for partial preferences of a customer under MNL model condition on her purchasing history and model parameters. As our empirical contribution, we quantify the predictive accuracy gains from introducing promotion effects in the partial order based nonparametric choice model and from accounting for consideration sets in a data-driven way using real-world panel data. Then we test the proposed methodology to optimize personalized promotions on the same sales transactions data.

One of the main goals of choice based demand estimation in general is to use the inferred model parameters as inputs in retail operations proposals, e.g., retailers’ revenue optimization. In the seminal work, van Ryzin and Mahajan (1999) propose a single-period stochastic inventory model, where customers substitute alternatives within the assortment while inventory is depleted. Most of the focus has been on the assortment problems in the context of rank-based choice models, e.g., papers by Jagabathula and Rusmevichientong (2016), Honhon et al. (2012) for retail operations assortments, and papers by Zhang and Cooper (2005), van Ryzin and Vulcano (2008), Kunnumkal (2014) for airline-related RM. We further refer a reader for a comprehensive review of assortment planning problems (Kök et al. 2008). We contribute to the stream of this literature in the context of retail operations assortment optimization. In particular, we formulate the personalized promotions algorithm to increase the retailers’ revenues as a mixed linear integer program (MILP) that can be efficiently solved. Consequently, we test our methodology to run personalized promotions on the real-world panel data (IRI academic dataset), resulting in considerable improvement of the retailer’s revenue.

The effect of sales promotions on a retailer is extensively studied in the field of marketing (Blattberg and Neslin 1990). A very illustrative overview of different promotions mechanisms used by a retailer is provided by Gedenk et al. (2006), where they summarize different objectives and effects of promotions known in the literature. Most of the empirical studies conclude that a sales promotion is a good instrument to induce the customer’s choice in the short run, however it has neutral or negative effect on the brand preference in the long run (Mela et al. 1997, Srinivasan et al. 2004, DelVecchio et al. 2006). There are very few papers reporting positive overall effect of sales promotions on brand preferences (Dekimpe et al. 1998, Foekens et al. 1998). In particular, DelVecchio et al. (2006) suggest that managers need to find a trade-off between the immediate boost in sales caused by drastic promotions and the long-run risk at which they place their brand by offering high price discounts. To find out if sales promotions can reduce the brand preference managers should also consider characteristics of their products, such as frequency of purchase, price level of the category, price level of the brand in the category, popularity of the brand, etc.. In addition, they claim that the propensity of promotions might be different for various customers.
Srinivasan et al. (2004) explore the question of whether sales promotions generate an additional revenue to manufacturer, retailer, or both. Their main result is that a price promotion does not have a permanent financial benefit neither for a retailer nor for a manufacturer. Winer (1986) studies the effect of the reference price (i.e., customer’s perception of the current price) on the probability of purchase of the brand. They argue that a lower price for an item today results not only in a higher current probability of purchase of the item but also negatively affects the long run sales by reducing the reference price. Therefore, if price promotions are very frequent, consumers may get used to lower prices and be reluctant to buy the item when the promotion is over and the price returns back to normal, i.e., it induces deal seeking behavior of customers. Raju (1995) reviews theoretical models of sales promotions. The author addresses the question of why firms implement sales mostly in mature markets and why price promotions are so popular in the retail industry. Also, the paper analyzes several models dealing with budget allocation between advertising and sales promotions, and customers’ response to the promotion. A vast majority of the literature in marketing assumes that customers’ consumption is fairly stationary. However, Ailawadi and Neslin (1998) investigate if consumption can be increased as the result of promotion sales. The authors believe that inventory level is the main channel of how promotions can influence the consumption, since price promotions usually result into a higher inventory. Higher inventory results in less stock-outs and higher usage rates. As a result, even if a company already has a high market share and can hardly attract new customers, it can still benefit from promotions by increasing consumer’s usage rates. Grover and Srinivasan (1992) determine different effects of promotions on different types of consumer segments. They divide consumers in different segments, for example, into brand loyal and brand switching segments by an iterative Bayesian procedure. The authors find that different segments are influenced by promotional variables differently. Van Heerde et al. (2003) argue that primary effects of promotions, like stockpiling and category expansion, are considerably higher than secondary effects, e.g., brand switching. As noted above, the focus in marketing community is typically on econometric modeling and estimation of massive (i.e., without personalization) promotions to derive managerial insights. Instead, one of the goals of the present paper is to develop a data-driven framework to run personalized promotions to optimize the revenue of the retailers by inducing brand switching effect. We believe that personalization is a key to mitigate the negative effects of price promotions (e.g., stockpiling) which dominate the outcome of massive promotions according to the marketing literature presented above.

Caro and Gallien (2012) present a clearance pricing optimization problem (i.e., relevant problem in fashion retailer industry during promotional period), which incorporates a demand forecasting model, as a linear integer program. More recently, Cohen et al. (2017b) formulate a linear integer programming approximation of the optimal single item promotion planning problem of grocery
Modeling the demand function is a crucial part of their analyses. Since the demand function is nonlinear in general, the authors use its linearization, assuming that demand’s variation can be explained by previous prices. In Cohen et al. (2017a), the authors extend this work for a setting when the retailer needs to decide the promotion discount for multiple items.

2. Choice Model Description

This section introduces a nonparametric choice-based demand model that is able to make fine-grained individual-level predictions while explicitly accounting for promotions. We will use this choice model later in Section 5 as an input for a personalized promotions algorithm to quantify customer response to price discounts. First, we describe the general modeling framework that is based on the partial orders of customers. Next, we provide the description of the data needed to calibrate our model. Then, we present the detailed explanation of the partial order construction process, and conclude with showing how to estimate our model.

2.1. Modeling Framework

Let us start with a description of the general modeling framework where we introduce some notation. We model preferences of customers with a general rank-based choice model of demand. We consider the universe of \( n \) unique substitutable products, where each item \( j \in \{1, 2, ..., n\} \) can be either under promotion or not. Consequently we represent every product \( j \in \{1, 2, ..., n\} \) either by its nonpromoted copy (or version) \( a_j \) or by its promoted copy (or version) \( a_{j+n} \). We define \( \mathcal{N} \) as the universe of nonpromoted copies of products, i.e., \( \mathcal{N} = \{a_1, a_2, ..., a_n\} \). Then we obtain \( \mathcal{N}' \) from \( \mathcal{N} \) by adding the promoted copies of all products, i.e., \( \mathcal{N}' = \{a_1, a_2, ..., a_{2n}\} \). In addition, let \( a_0 \) denote a “no purchase” or “outside option” which is assumed to be always available for customers.

Preferences over the universe of product copies \( \mathcal{N}' \cup \{a_0\} \) are represented by an antireflective, antisymmetric, and transitive relation \( \succ \), which induces a total ordering or ranking over all the copies of products, and we use the expression \( a_k \succ a_j \) to say that \( a_k \) is preferred to \( a_j \). Equivalently, the preference relations can be captured through rankings or permutations. Every preference list \( \sigma \) represents a ranking over \( 2n + 1 \) items in the universe \( \mathcal{N}' \cup \{a_0\} \), where \( \sigma(a_j) \) denotes the preference rank of product \( a_j \). The lower rankings indicate the higher preference order. Therefore, for customers described by preference list \( \sigma \) we have that \( a_k \succ \sigma a_j \).

The population consists of \( m \) customers making repeated purchases over \( T \) discrete time periods. The set of customers and products is assumed to remain the same over time. Every customer \( i \in \{1, 2, ..., m\} \) is represented by a directed acyclic graph (DAG) \( D_i \), or equivalently, by a partial order (PO) \( D_i \). A general partial order \( D_i \) specifies a collection of edges of DAG \( D_i \) (or a collection
of pairwise preference relations) denoted by $E_{D_i} \subset \{(a_k, a_j) : 0 \leq k, j \leq 2n, k \neq j\}$, so that for any $(a_k, a_j) \in E_{D_i}$ we have that item $a_k$ is preferred to item $a_j$. In every store visit, customer $i$ samples a full ranking of products consistent with her DAG $D_i$ from a distribution denoted by $\lambda$. We say that a full ranking $\sigma$ is consistent with partial order $D_i$ if and only if for every $(a_k, a_j) \in E_{D_i}$ we have that $\sigma(a_k) < \sigma(a_j)$.

In a nutshell, the choice process can be described as follows. Upon visiting the store, customers are offered a subset of products $S$ from the universe $\mathcal{N}'$. Next each customer $i$ samples her preference list $\sigma$ according to distribution $\lambda$ from the collection of full rankings consistent with partial order $D_i$. Then according to the sampled ranking $\sigma$ she purchases the most preferred product $a_k$ within the set of considered items $C$, which is a subset of the offer set $S$, i.e., $a_k = \arg \min_{a_i \in C} \sigma(a_i)$.

### 2.2. Model Assumptions

Suppose we have a panel data of $m$ customers making repeated purchases over $T$ periods. Every customer $i$ makes $T_i$ purchases over the specified time horizon $T$. To simplify the notation we relabel the periods on a per customer basis, so that $t=1$ corresponds to customer’s first visit to the store, $t=2$ corresponds to customer’s second visit, and so on until the last purchasing instance at time $T_i$. Historical purchase information of each customer $i$ can be stored as the number of tuples $(a_{j_{it}}, S_{it})$ for $t=1, 2, ..., T_i$, where $S_{it} \subset \mathcal{N}'$ is the set of products offered to a customer $i$ at time $t$, and $a_{j_{it}} \in S_{it}$ is the product purchased by a customer $i$ at time $t$.

Note that every tuple implicitly contains information about product promotions. Based on the specified notation, if $a_j \in S_{it}$ and $j \geq n+1$ then $a_j$ is a promoted copy of product $j - n \in \{1, 2, ..., n\}$, i.e., product $j - n$ is offered to an individual $i$ under promotion at time $t$. Similarly, if $a_j \in S_{it}$ and $j < n+1$ then $a_j$ is a non-promoted copy of product $j \in \{1, 2, ..., n\}$, i.e., product $j$ is offered to an individual $i$, and it is not under promotion at time $t$. Each product can be either under display promotion (i.e., product is prominently displayed on the store shelves) or under price reduction promotion (i.e., product is offered at a discounted price). To simplify our analysis we assume that the promotion feature is a binary attribute of a product and we restrict ourselves to only price reduction promotions. However, our model can be extended to account for several types of promotions (e.g., price reduction and display promotions) and finite amount of depths within each type of promotions by further expanding the product universe.

### 2.3. Partial Order Construction: Promotion Model

The potential benefit of the DAG-based representation of individual preferences is the boost in forecasting accuracy it could provide to make fine-grained individual-level purchase predictions. These gains could be particularly significant for cases in which only a small sample of information
is available for each customer. To infer the signal from the limited data efficiently, we model each individual through a partial order which consists of two copies of each product: when the product is under promotion and otherwise. Specifically, the DAG $D$ for each customer $i$ is constructed in the following three phases:

**Phase 0: Initializing a Preference Graph and Adding Strong Edges.** We start from empty graph $G$, and add $2n$ isolated nodes from the product universe $N'$, where each node represents a product either under promoted or full price. Then we add additional node $a_0$ which corresponds to the no purchase option. Let $E_G$ denote the set of edges in the graph $G$. Next for every item $j \in \{1, \ldots, n\}$ we add a strong edge to the graph $G$ from its promoted copy $a_{j+n}$ to its nonpromoted copy $a_j$, i.e., $E_G \leftarrow E_G \cup \{(a_{j+n}, a_j) : \forall j \in \{1, \ldots, n\}\}$. It is assumed that promoted copy $a_{j+n}$ of every product $j \in \{1, \ldots, n\}$ is preferred to its nonpromoted copy $a_j$ since both products have the same attributes except the price which is lower when a product is under promotion.

**Phase 1: Adding Candidate Edges from Sales Transactions.** We add candidate edges to the preference graph $G$ keeping track of the sequence of customers’ visits to the store obtained from the panel data. In particular, we sequentially add candidate arcs (pairwise comparisons) to the graph $G$ inferred from the sequence of purchases of a customer. For each purchasing instance of an individual $i$ that can be described by tuple $(a_{j_{it}}, S_{it})$ we draw edges from $a_{j_{it}}$ to items in $S_{it} \cup \{a_0\} \setminus \{a_{j_{it}}\}$, i.e., $E_G \leftarrow E_G \cup \{(a_{j_{it}}, a_\ell) : \forall a_\ell \in S_{it} \cup \{a_0\} \setminus \{a_{j_{it}}\}\}$. We add those edges since it follows from our model that the customer purchases the product $a_{j_{it}}$ if $a_{j_{it}} = \arg \min_{a_\ell \in S_{it} \cup \{a_0\}} \sigma(a_\ell)$, so that $\sigma(a_{j_{it}}) < \sigma(a_\ell)$ for all $a_\ell \in S_{it} \cup \{a_0\} \setminus \{a_{j_{it}}\}$, where $\sigma$ is the preference list that was sampled by a customer to make a purchase. Some of the candidate edges that we add to the preference graph during this phase might be spurious (i.e, erroneous) since some of the customers might not consider all the products in the offer set before making a purchase. Therefore in the next phase we are going to infer (i.e, identify) spurious edges in the preference graph in a data-driven way and delete them. In addition, we assign the weight $w_{j,\ell}$ (i.e., the counter of the candidate edge $(a_j, a_\ell)$) for every candidate edge that we add during this phase, such that $w_{j,\ell}$ is the number of times the product $a_j$ was preferred to $a_\ell$ (or the number of times the candidate edge $(a_j, a_\ell)$ shows up in the data).

**Phase 2: Preference Graph Decycling while Accounting for Consideration Sets.** A preference graph obtained after the previous phase can have spurious candidate edges because we ignored the consideration set formation of customers in Phase 1. Consequently, these spurious candidate edges might result into the cycles in the preference graph of a customer. Accounting for consideration sets of a customer in a data-driven way, we decycle the preference graph of individuals by identifying spurious candidate edges and deleting them. In order to formalize the process of inferring a partial order through preference graph decycling (or equivalently, finding spurious edges
in the preference graph and deleting them) we formulate an optimization problem that needs to be solved for every customer with cycles in the preference graph, where we maximize the expected size of the consideration set of each store visit, such that Phases 0 and 1 result in a preference graph without cycles. Let \( C(a_{j,t}, S_{it}) \) be a star graph with a root \( a_{j,t} \) and leaves \( S_{it} \cup \{a_0\} \backslash \{a_{j,t}\} \) that corresponds to the purchasing transaction of individual \( i \) at time \( t \). Let \( E_G \) be the set of edges in graph \( G \). Note that every candidate edge \((a_\ell, a_j)\) in the graph \( G \) has a weight \( w_{\ell,j} \). Let us define a binary variable \( x_{k\ell} \) for all \((a_k, a_\ell)\in E_G\) as follows:

\[
x_{k\ell} = \begin{cases} 
1, & \text{if } (a_k, a_\ell)\in E_D, \\
0, & \text{otherwise},
\end{cases}
\]

where \( D \) is the partial order that is obtained from the preference graph \( G \) by deleting its spurious edges. The optimization problem of an individual \( i \) to infer her partial order \( D \) is given by

\[
\max_x \sum_{t=1}^{T_i} \sum_{a_k\in S_{it}} x_{j,t,k} \\
\text{s.t.: } D \text{ is a Partial Order,} \\
x_{k\ell} = 1, \quad \forall (a_k, a_\ell)\in E_G, \text{ if } (a_k, a_\ell) \text{ is a strong edge.}
\]
where we maximize the expected consideration set size of an individual \( i \) such that \( D \) is a DAG. It is clear that the objective function of the optimization problem above can be rewritten as
\[
\max_x \sum_{(a_k, a_\ell) \in E'_G} w_{k\ell},
\]
s.t.: \( b_{k\ell} \geq x_{k\ell}, \quad \forall (a_k, a_\ell) \in E_G, \)
\[
b_{k\ell} + b_{\ell k} = 1, \quad \forall a_k, a_\ell \in \mathcal{N}', \quad k \leq \ell,
\]
\[
b_{k\ell} + b_{\ell p} + b_{pk} \leq 2, \quad \forall a_k, a_\ell, a_p \in \mathcal{N}', \quad k \neq \ell \neq p,
\]
\[
x_{k\ell} = 1, \quad \forall (a_k, a_\ell) \in E_G, \quad \text{if } (a_k, a_\ell) \text{ is a strong edge},
\]
\[
b_{k\ell} \in \{0, 1\}, \quad \forall a_k, a_\ell \in \mathcal{N}', \quad k \neq \ell,
\]
\[
x_{k\ell} \in \{0, 1\}, \quad \forall (a_k, a_\ell) \in E_G,
\]

where the constraints of the optimization problem above ensure that \( D \), which is a subgraph of \( G \) induced by the decision variables \( x_{k\ell} \), is a partial order. In particular, the second set of equalities ensures that either product \( a_k \) goes before product \( a_\ell \) in the partial order, or product \( a_\ell \) goes before product \( a_k \). The third set of constraints ensure a linear ordering among three products. The first set of constraints together with the second set of equalities and the third set of constraints ensure that there will be no cycles in \( D \), i.e., in the subgraph of \( G \) characterized by the decision variables \( x_{k\ell} \). Note that the last binary constraint can be relaxed by the linear constraint \( 0 \leq x_{k\ell} \leq 1, \quad \forall (a_k, a_\ell) \in E_G \), because \( b_{k\ell} \) is a binary variable, so that we can consider only two cases: (1) if \( b_{k\ell} = 1 \), then in combination with the linear constraint \( x_{k\ell} \) we have that \( x_{k\ell} = 1 \) to maximize the objective function; (2) if \( b_{k\ell} = 0 \), then in combination with the linear constraint \( x_{k\ell} \geq 0 \) we have that \( x_{k\ell} = 0 \). Since solving the MIP to optimality is challenging in general we develop a heuristic algorithm for preference graph decycling (see Section A3.1 in the Appendix) and show on actual sales transaction data that this algorithm performs well in practice.

**Phase 3: Adding Implicit Candidate Edges to the Inferred DAG.** Finally, we add implicit edges to the DAG \( D \) that can be inferred from the DAG \( D \) itself. First, let us relabel all the edges in the DAG \( D \), obtained after Phase 2, as *strong*, i.e., from now on we assume that all the edges in the DAG \( D \) obtained after the previous phase are *strong*. Then for any edge \((a_\ell, a_j) \in E_D\), if the edge \((a_\ell, a_j)\) is not already in \( D \) we add a corresponding candidate edge \((a_\ell, a_j)\) to \( D \), i.e., \( E_D \leftarrow E_D \cup (a_\ell, a_j) \), and then we assign the weight \( w_{\ell j} \) to the candidate edge \((a_\ell, a_j)\) of \( D \) s.t. \( w_{\ell j} = \).
Similarly, for any edge \((a_\ell, a_j) \in E_D\) if the edge \((a_\ell+n, a_j+n)\) is not already in \(D\) we add a corresponding candidate edge \((a_\ell+n, a_j+n)\) to \(D\), i.e., \(E_D \leftarrow E_D \cup (a_\ell+n, a_j+n)\), and then we assign the weight \(w_{\ell+n,j+n}\) to the candidate edge \((a_\ell+n, a_j+n)\) of \(D\) such that \(w_{\ell+n,j+n} = w_{\ell,j}\). By adding these implicit candidate edges, we assume that if product \(\ell\) is preferred to product \(j\) when both of them are under full price (i.e., \((a_\ell, a_j) \in E_D\)), then product \(\ell\) is likely to be preferred to product \(j\) when both of them are under promotion (i.e., \((a_\ell+n, a_j+n) \in E_D\)). Similarly, if product \(\ell\) is preferred to product \(j\) when both of them are under promotion, then product \(\ell\) is likely to be preferred to product \(j\) when both of them are under full price. To this end if adding implicit candidate edges results in cycles in \(D\) we infer the partial order of a customer by renaming preference graph \(D\) (which consists of strong and candidate edges) by \(G\) and applying the preference decycling optimization problem (2) to \(G\).

2.3.1. Illustrative Example of DAG Construction. Figure 1 illustrates the construction of a DAG \(D\) for an individual \(i\) in four phases. The horizon has \(T_i = 3\) periods, and there are \(n = 4\) products in the category. For every product \(j \in \{1, 2, 3, 4\}\) we have a promoted copy \(j''\) (i.e., \(j''\) is equivalent to copy \(a_{j+n}\)) and a nonpromoted copy \(j'\) (i.e., \(j'\) is equivalent to copy \(a_j\)), where we rename product copies for ease of exposition. In Phase 0, starting from the empty preference graph, we add no purchase option 0 and two copies for every product \(j \in \{1, 2, 3, 4\}\) (e.g., add nodes 1' and 1''). Then we add strong edges from promoted copy of every product to its nonpromoted copy (e.g., add a strong edge from 1'' to 1'). In Phase 1, for every purchasing transaction we draw edges from purchased item to other items in the offer set, ignoring consideration set formation of customers (e.g., at time \(t = 1\) we draw edges from purchased item 2' to other items in the offer set: 1', 3'', and 0). Note that Phase 1 results in a cycle between copies 2' and 3'' in the preference graph. As a result, one of the edges between 2' and 3'' is spurious, with \(w_{2'3''} = 3\) and \(w_{3''2'} = 3\). In Phase 2, maximizing the expected consideration set size of an individual, we find that the edge \((3'',2')\) is spurious and delete it from the preference graph, i.e., consideration set at time \(t = 2\) is smaller than the offer set and consists only of the items 3'' and 4'. Phase 3 concludes the DAG construction process where we add to the DAG \(D\) an implicit edge \((2'',1'')\), which corresponds to the edge \((2',1')\) in the DAG obtained after Phase 2. In this case there are no new cycles, but if they were, we run Phase 2.

2.4. MLE of Partial Order based Choice Model

Once we infer the partial orders of customers following the preference graph construction process described above (see Section 2.3), we use the Maximum Likelihood Estimation (MLE) framework to calibrate our model. Defining \(\lambda(\sigma)\) as the probability assigned to a full ranking \(\sigma\), we can interpret
any partial order $D$ as a censored representation of underlying full ranking $\sigma$ that a customer samples according to the distribution $\lambda$ over the full rankings. Let $S_D$ denote the set of compatible with $D$ permutations of products’ full rankings, i.e., $S_D = \{\sigma : \sigma(a_i) < \sigma(a_j) \text{ whenever } (a_i, a_j) \in D\}$. As a result, we can compute the likelihood of the DAG $D$ as follows:

$$\lambda(D) = \sum_{\sigma \in S_D} \lambda(\sigma).$$

Then the panel data log-likelihood function is given by

$$\log L(\text{Panel Data} \mid \beta) \Delta= \sum_{i=1}^{m} \log \Pr[(a_{jit}, S_{it})_{t=1}^{T_i} \mid D_i, \beta]$$

$$= \sum_{i=1}^{m} \log \left( \Pr(D_i \mid \beta) \cdot \Pr[(a_{jit}, S_{it})_{t=1}^{T_i} \mid \beta, D_i] \right)$$

$$= \sum_{i=1}^{m} \log \left( \Pr(D_i \mid \beta) \cdot \prod_{t=1}^{T_i} \Pr[(a_{jit}, S_{it}) \mid \beta, D_i] \right)$$

$$= \sum_{i=1}^{m} \log \Pr(D_i \mid \beta) + \sum_{i=1}^{m} \sum_{t=1}^{T_i} \log f(a_{jit}, S_{it}, D_i)$$

$$= \sum_{i=1}^{m} \log \Pr(D_i \mid \beta) = \sum_{i=1}^{m} \log \lambda(D_i) = \sum_{i=1}^{m} \log \left( \sum_{\sigma \in S_{D_i}} \lambda(\sigma) \right),$$

where $\beta_j$ is the nominal utility of item $a_j \in N'$ and $f(a_{jit}, S_{it}, D_i)$ is the probability to purchase item $a_{jit}$ from the offer set $S_{it}$ for individual $i$ with DAG $D_i$. The second equality follows from a straightforward application of a conditional probability formula. The third equality follows because conditioning on the DAG $D_i$, individual $i$’s purchase probabilities can be computed independently. The fifth equality follows since $a_{jit}$ is a top ranked product within the items in $S_{it}$ according to all the preference lists that are consistent with DAG $D_i$, $\forall i \in \{1, ..., m\}$ and $\forall t \in \{1, ..., T_i\}$. As a result, the likelihood function that we maximize to calibrate the model is the sum of likelihoods of customers’ partial orders. Note that the tractability of the log-likelihood function depends on the distribution $\lambda$ over preference lists, e.g., it is concave under MNL/Plackett-Luce distribution of $\lambda$.

### 3. Theoretical analyses of PO-MNL Model

We can potentially define different probability distributions over the set of rankings (e.g., see Marden 1995), however, our focus in the present paper is on the specification of such probabilities under the Plackett Luce (PL) model. It follows from the equation (3) that computing the likelihood of a general partial order $D$ is #P-hard problem, because of the #P-hardness of counting the number of full rankings consistent with $D$ (Birghtwell and Winkler 1991). First, we consider $\tilde{\lambda}(D)$,
i.e., the tractable approximation of the likelihood of a DAG \( D \) proposed by Jagabathula and Vulcano (2017):

\[
\tilde{\lambda}(D) = \prod_{a_j \in \mathcal{N}} \frac{v_j}{\sum_{a_k \in \Psi_D(a_j)} v_k},
\]

where \( v_j = \exp(\beta_j) \), \( \forall a_j \in \mathcal{N} \), for nominal utility \( \beta_j \) of product \( j \), and \( \Psi_D(a_j) \) denotes the reachability function such that \( \Psi_D(a_j) = \{a_k : a_k \text{ is reachable from } a_j \text{ in } D\} \). Note that \( \Psi_D(a_j) \) is always nonempty, since we assume that each node \( a_j \) is reachable from itself. The approximation of the likelihood of DAG \( D \) in equation (4) is exact when \( D \) is a forest of directed trees, each with a unique root. Then in the paper by Jagabathula and Vulcano (2017) the approximate probability to choose product \( a_j \) from the offer set \( S \) when a preference list is sampled consistently with DAG \( D \) is given by

\[
\tilde{f}(a_j, S, D) = \begin{cases} 
\frac{v_{\Psi_D(a_j)}}{\sum_{a_k \in h_D(S) \Psi_D(a_k)}} & \text{if } a_j \in h_D(S), \\
0 & \text{otherwise}, 
\end{cases}
\]

where \( h_D(S) \) are roots of the subgraph induced in \( D \) by the offer set \( S \), and \( v_A = \sum_{a_j \in A} v_j \). These probabilities are exact when DAG \( D \) is a forest of directed trees each with a unique head and when the heads of the subgraph induced in \( D \) by the offer set \( S \) are also heads of \( D \).

### 3.1. Tractable Analytical Bounds on the Likelihood of Partial Order

We can easily compute the transitive reduction for an arbitrary DAG \( D \) (Aho et al. 1972), i.e., an equivalent DAG that has as few edges as possible having the same reachability relationship as \( D \). In further analyses we assume that we operate with a transitive reduction of DAG \( D \). Let \( \nu \)-node denote a node in DAG \( D \) that has more than one incoming edge and let \( \mathcal{F}_D \) denote the set of \( \nu \)-nodes in DAG \( D \). The presence of \( \nu \)-nodes in the DAG \( D \) makes \#P-hard to compute its likelihood. Therefore, our goal in this subsection is to find computationally tractable upper and lower bounds on the likelihood of a DAG with \( \nu \)-nodes. In the next proposition we show that \( \tilde{\lambda}(D) \) obtained from equation (4) is a lower bound of the DAG \( D \)'s likelihood.

**Proposition 1.** Suppose \( D \) is a DAG. Then under Placket Luce model

\[
\tilde{\lambda}(D) \leq \lambda(D),
\]

with strict inequality if all the parameters are strictly positive, and there is at least one \( \nu \)-node in \( D \).

Next in order to find the upper bound approximation of DAG \( D \)'s likelihood, let us denote \( \bar{D} \) as the DAG obtained from \( D \) where for every node with more than one incoming edge in DAG \( D \) we delete all the incoming edges, except exactly one (e.g., \( \bar{D} = \Phi(D) \), see Section A3.2 in the Appendix). Then the next proposition follows.
Proposition 2. Suppose $D$ is a DAG, then
\[
\lambda(D) \leq \lambda(\bar{D})
\]
with strict inequality under Plackett Luce model if all the parameter values are strictly positive, and there is at least one $v$-node in $D$.

The Proposition 2 follows from Lemma A3 in the Appendix, since $\bar{D}$ is obtained from $D$ by deleting some of its edges. Next we explore the tightness of the developed lower and upper bounds of a DAG’s likelihood. Consequently, let $R(D)$ denote the tightness of the bounds such that $R(D) = \frac{\lambda(\bar{D})}{\lambda(D)}$ and let $V_D$ denote the set of nodes in a DAG $D$. Note that $R(D) \geq 1$ since it follows from Propositions 1 and 2 that $\tilde{\lambda}(D) \leq \lambda(D) \leq \lambda(\bar{D})$. Let $\ell$ denote the maximum size of the reachability set in $D$, i.e.,
\[
\ell = \max \limits_{a_j \in V_D} |\Psi_D(a_j)|
\]
and let $p$ denote the number of nodes in $D$ that have $v$-nodes in their reachability set, i.e.,
\[
p = \sum \limits_{a_j \in V_D} [\mathbb{I}\{\mathcal{F}_D \cap \Psi_D(a_j) \neq \emptyset\}],
\]
where $\mathbb{I}[A]$ is one if a condition $A$ is satisfied and zero otherwise. Then the next proposition follows.

Proposition 3. Suppose $\Delta = \max \frac{\sum \limits_{k \neq j} a_k}{v_j}$ s.t. $k \neq j$, $a_k \in \Psi_D(a_j)$, and $\mathcal{F}_D \cap \Psi_D(a_j) \neq \emptyset$, then
\[
0 \leq \log R(D) \leq p \cdot \log(1 + \ell \cdot \Delta),
\]
then assuming $n$ is the number of nodes in $D$, i.e., $n = |V_D|$, and $\Delta = o(n^{-2})$ we have that
\[
\lim \limits_{n \to \infty} \tilde{\lambda}(D) = \lambda(D),
\]
since
\[
0 \leq \lim \limits_{n \to \infty} \log R(D) \leq \lim \limits_{n \to \infty} \Delta \cdot n^2.
\]

After we characterize the analytical upper and lower bounds of a DAG’s likelihood as a function of the model parameters, we show on the actual sales data that these bounds perform well in practice (see Section A3.2 in the Appendix).

3.2. Tractable Analytical Bounds on the Purchase Prediction

Denoting $D_1 \cup D_2$ as the merge (or the union) of graphs $D_1$ and $D_2$, we consider $f(a_j, S, D)$, i.e., the exact probability of choosing product $a_j$ from the offer set $S$ when a preference list is sampled consistently with DAG $D$:
\[
f(a_j, S, D) \triangleq \Pr(S_{C(a_j, S)}|S_D) = \frac{\Pr(S_{C(a_j, S)} \cap S_D)}{\Pr(S_D)} = \begin{cases} \frac{\lambda(D \cup C(a_j, S))}{\lambda(D)}, & \text{if } a_j \in h_D(S), \\ 0, & \text{otherwise}, \end{cases}
\]
where the second equality is obtained from the straightforward application of the Bayes rule. The last equality in the equation above is valid since any product $a_j \notin h_D(S)$ has an incoming edge from another product in the offer set $S$, so that item $a_j$ is not going to be purchased if a customer makes
purchases consistently with her partial order $D$ (i.e., according to every preference list consistent with the DAG product $a_j$ is less preferred than another product in the offer set). Next, we define an approximate probability (since it is a #P-hard problem to calculate the exact probabilities) of choosing product $a_j$ from the offer set $S$ when preference list is sampled consistently with DAG $D$, providing also upper and lower bounds of these probabilities.

**Definition 1.** Suppose we are given a DAG $D$. Then, under Plackett-Luce model, the approximate probability of choosing product $a_j$ from the offer set $S$ assuming that the sampled preference list is consistent with DAG $D$ is given by

$$
\hat{f}(a_j, S, D) = \begin{cases} 
\frac{\lambda(D \cup C(a_j, S))}{\lambda(D)}, & \text{if } a_j \in h_D(S), \\
0, & \text{otherwise.}
\end{cases}
$$

Consequently, let $f(a_j, S, D)$ denote the lower bound of the purchase probability, such that $f(a_j, S, D) = \frac{\lambda(D \cup C(a_j, S))}{\lambda(D)}$ if $a_j \in h_D(S)$ and 0, otherwise. Let $\bar{f}(a_j, S, D)$ denote the upper bound of the purchase probability, such that $\bar{f}(a_j, S, D) = \frac{\lambda(D \cup C(a_j, S))}{\lambda(D)}$ if $a_j \in h_D(S)$, and 0, otherwise. Now the corollary follows.

**Corollary 1.** Suppose we are given a DAG $D$. Then, under Plackett-Luce model, the following tractable bounds of purchase probabilities apply:

$$
f(a_j, S, D) \leq \hat{f}(a_j, S, D), \hat{f}(a_j, S, D) \leq \bar{f}(a_j, S, D).
$$

In the next proposition we analyze the tightness of the proposed lower and upper bounds $\underline{f}(a_j, S, D)$ and $\bar{f}(a_j, S, D)$. Let $C_j$ be a star graph with directed edges from the root $a_j$ to other nodes in the offer set $S$, i.e., $C_j = C(a_j, S)$. Then let $\ell$ denote the maximum size of the reachability set in the DAG $D \cup C_j$, i.e., $\ell = \max_{a_k \in \Psi_{D \cup C_j}(a_j)} |\Psi_{D \cup C_j}(a_k)|$ and let $p$ denote the number of nodes in $D \cup C_j$ that have $v$-nodes in their reachability set, i.e., $p = \sum_{a_m \in V_{D \cup C_j}} |\Psi_{D \cup C_j}(a_m) \neq \emptyset|$, where $|A|$ is one if a condition $A$ is satisfied and zero otherwise. Then the proposition below follows.

**Proposition 4.** Suppose $\Delta = \max_{v \in m} s.t. k \neq m$, $a_k \in \Psi_{D \cup C_j}(a_m)$, and $\Psi_{D \cup C_j}(a_m) \neq \emptyset$, then

$$
0 \leq \log \frac{\bar{f}(a_j, S, D)}{\underline{f}(a_j, S, D)} \leq 2 \cdot p \cdot \log(1 + \ell \cdot \Delta),
$$

then assuming $n$ is the number of nodes in $D$, i.e., $n = |V_D|$, and $\Delta = o(n^{-2})$ we have that

$$
\lim_{n \to \infty} \hat{f}(a_j, S, D) = f(a_j, S, D),
$$

since

$$
0 \leq \lim_{n \to \infty} \log \frac{\bar{f}(a_j, S, D)}{\underline{f}(a_j, S, D)} \leq \lim_{n \to \infty} 2 \cdot \Delta \cdot n^2.
$$

After providing the analytical upper and lower bounds on the prediction guarantees for the partial order based choice model, we show on actual sales data that these tractable bounds perform well in practice (see Section A3.2 in the Appendix).
4. Empirical Study

In this section, we present the empirical results using the IRI Academic dataset (Bronnenberg et al. 2008), which consists of real-world purchase transactions from grocery and drug stores. We start describing the data that we use to calibrate and test our model. Next we specify the models that we compare. Then we define prediction performance measures that we use to test the models. Finally, we fit the PO-MNL Promotion model in a horse-race against two state-of-the-art benchmarks, the latent class MNL (LC-MNL) and the Random Parameters Logit (RPL) models. We compare all the choice models on the accuracy of two prediction measures on hold-out data, and find that PO-MNL Promotion model outperforms the benchmarks in most of the categories analyzed.

4.1. Data Analysis

We analyze one year (calendar year 2007) of consumer packaged goods (CPG) purchase transaction data for chains of grocery stores in the two largest Behavior Scan markets. For every purchase instance in the data set, we have the week and the store id of the purchase, the Universal Product Code (UPC) of the purchased item, the panel id of the purchasing customer, quantity purchased, price paid, and an indicator of whether the purchased item is on promotion. Overall we considered 27 categories (see Table 1) of goods out of 31 ignoring “photography supplies”, “razors”, “blades”, and “diapers” because of the data sparsity. The data consists of 1.2M records of weekly purchase transactions from 84K customers over 52 weeks. The transaction data is split into two parts: the training set that consists of the first 26 weeks of purchase observations, and the test set that consists of the last 26 weeks of purchase observations. We consider only customers with two or more transactions over the training period. In total, we have 64K customers and 1.1M purchase transactions. To alleviate data sparsity the sales data are aggregated by vendor in the following way. We aggregate all the products with the same vendor code (comprising digits 3 through 7 in 13-digit-long UPC code) into a unique product. The reader is referred to Table 1 for a detailed summary of the training data.

Based on Table 1 we have 46.4% of individuals that have preference graphs without cycles under under PO Promotion model. Note that individuals with cycles in their preference graph have a denser DAG than individuals without cycles in their preference graph for all the categories of products. In particular, individuals with cycles in their preference graph on average (on average across 27 categories of products) have 41.9% more edges in their DAG. Across 27 categories of products, on average we have 50 vendors per category, and the average assortment consists of 35.2% of the total category vendors.
### Table 1 Description of the data. The column 'Vend' - the number of products obtained in the training data after aggregating different UPCs by vendor; 'AvOS' - the average # of vendors in the offer set; 'Total'-the total # of individuals in the training data; '≥2 sales' - # of individuals with at least two purchases in the training data; 'AvTr' - the average # of transactions per customer in the training data. Analyzing customers under PO Promotion model, we report 'NoC' - # of individuals without cycles in the preference graph after Phase 1; 'Dens1' - the average # of edges in the DAG for individuals without cycles in the preference graph after Phase 1; 'Dens2' - the average # of edges in the DAG for individuals with cycles in the preference graph after Phase 1.

#### 4.2. Models Compared

Having the sales data described above, we calibrate the PO-MNL Promotion model (proposed in this paper to measure customer response to product promotions) against two widely used benchmarks: the latent class MNL (LC-MNL) and the Random Parameters Logit (RPL) models. All
these models belong to the class of general random utility maximization (RUM) choice rule. Therefore, in each purchase instance every customer samples product utilities $\beta_j$, and then chooses the product giving the highest one.

4.2.1. LC PO-MNL Promotion Model First, we consider a single class PO-MNL Promotion model. Let $\tau_j$ denote a parameter value of any product copy $a_j \in N^i$ s.t.

$$
\tau_j = \begin{cases} 
\beta_j^0, & \text{if } 0 \leq j \leq n \\
\beta_{j-n}^0 + \beta_j, & \text{if } n+1 \leq j \leq 2n,
\end{cases}
$$

where $\forall j \in \{0,1,\ldots,n\}: \beta_j^0$ is the parameter value of product $j$ under full price, and $\forall j \in \{n+1,n+2,\ldots,2n\}: \beta_{j-n}^0 + \beta_j$ is the parameter value of product $j-n$ under promoted price. Then we estimate parameters of PO-MNL Promotion model by solving the following approximated regularized maximum likelihood estimation problem

$$
\max_{\beta,\beta_0} \sum_{i=1}^{m} \sum_{j=1}^{2n} \left[ \tau_j - \log \left( \sum_{a_j \in \Psi_{D_i}(a_j)} \exp(\tau_j) \right) \right] - \alpha(\|\beta^0\|_1 + \|\beta\|_1),
$$

where $\Psi_{D_i}(a_j)$ is the set of nodes that are reachable from $a_j$ in DAG $D_i$ of customer $i$. When the value of $\alpha$ is fixed, it can be shown that the optimization problem in (6) is globally concave and therefore can be solved efficiently (Train 2009). We turned the value of $\alpha$ by 5-fold cross-validation.

The likelihood function is exact only if every DAG $D_i$ is a forest of directed trees each with a unique head. Then, the approximate probability to choose product $a_j$ from the offer set $S$ when preference list is consistent with a DAG $D$ can be written as

$$
\tilde{f}(a_j, S, D) = \begin{cases} 
\frac{\sum_{a_j \in \Psi_{D_i}(a_j)} \exp(\tau_j)}{\sum_{a_j \in h_{D}(S)} \sum_{a_j \in \Psi_{D_i}(a_j)} \exp(\tau_j)}, & \text{if } a_j \in h_{D}(S), \\
0, & \text{otherwise},
\end{cases}
$$

where $h_{D}(S)$ are roots of the subgraph induced in $D$ by the offer set $S$, and $v_A = \sum_{a_j \in A} \exp(\tau_j)$. These probabilities are exact when a customer’s DAG $D$ is a forest of directed trees each with a unique head and the heads of the subgraph induced in $D$ by the offer set $S$ are also heads of $D$.

Then, we consider a $K$-latent-class PO-MNL Promotion model, where it is assumed that each customer belongs to one of the $h \in \{1,\ldots,K\}$ latent classes. An individual from the class $h$ samples her DAG in accordance with PO-MNL Promotion model where the parameter value of any product copy $a_j \in N^i$ is defined by $\tau_{jh}$ s.t.

$$
\tau_{jh} = \begin{cases} 
\beta_{j-h}^0, & \text{if } 0 \leq j \leq n \\
\beta_{j-n-h}^0 + \beta_{j-n-h}, & \text{if } n+1 \leq j \leq 2n.
\end{cases}
$$

A prior probability of a customer to belong to the class $h$ is $\gamma_h \geq 0$ such that $\sum_{h=1}^{K} \gamma_h = 1$. Then similar to the PO-MNL model the regularized maximum likelihood problem can be formulated as follows:

$$
\max_{\beta,\beta_0,\gamma} \sum_{i=1}^{m} \log \left( \sum_{h=1}^{K} \gamma_h \prod_{j=1}^{2n} \frac{\exp(\tau_{jh})}{\sum_{a_j \in \Psi_{D_i}(a_j)} \exp(\tau_{jh})} \right) - \alpha \sum_{h=1}^{K} (\|\beta^0_h\|_1 + \|\beta_h\|_1),
$$
where $\Psi_{D_i}(a_j)$ is the set of nodes that can be reached from $a_j$ in DAG $D_i$ of a customer $i$. Since the above optimization problem is nonconcave for $K > 1$ even for a fixed $\alpha$, we use the expectation-maximization method (EM) described in Train (2009) to estimate the parameters (see the details in Appendix A2.1.2 in Jagabathula and Vulcano 2017). Specifically, we initialize the EM with a random allocation of customers to one of the $K$ classes, resulting in an initial allocation $D_1, D_2, \ldots, D_K$, which form a partition of the collection of all the customers. Then we set $\hat{\gamma}_h^{(0)} = |D_h|/(\sum_{d=1}^K |D_d|)$. In order to get a parameter vector $\tau_h^{(0)}$, we fit a PO-MNL Promotion model, described above, to each subset of customers. Then, using $\{\gamma^{(0)}, (\tau_h^{(0)})_{h=1}^K\}$ as the starting point, we carried out EM iterations.

Given the parameter estimates, we make predictions as follows. For each individual $i$ with DAG $D_i$, we estimate the posterior membership probabilities $\hat{\gamma}_{ih}$ for each class $h \in [1, \ldots, K]$:

$$
\hat{\gamma}_{ih} = \frac{\gamma_h \prod_{j=1}^{2n} \left[ v_{jh} / \sum_{a_l \in \Psi_{D_i}(a_j)} v_{hl} \right]}{\sum_{d=1}^K \gamma_d \prod_{j=1}^{2n} \left[ v_{jd} / \sum_{a_l \in \Psi_{D_i}(a_j)} v_{ld} \right]},
$$

where $v_{jh} = \exp(\tau_{jh})$, and then make predictions in the following way:

$$
\tilde{f}(a_j, S, D_i) = \sum_{h=1}^K \hat{\gamma}_{ih} \tilde{f}_h(a_j, S, D),
$$

where $\tilde{f}_h(a_j, S, D)$ are approximated probabilities from PO-MNL Promotion Model.

4.2.2. Benchmark Models  We compare our models with two benchmarks described here. The first benchmark is the $K$ latent class MNL (LC-MNL) choice model (see Appendix A2 for details), where each customer belongs to one of $h \in \{1, \ldots, K\}$ unobservable classes, and remains there during the whole purchase horizon. Purchasing transactions are considered to be independent realizations keeping track of the individual id in the panel data. According to this model, in order to make individual-level predictions we average the predictions from $K$ single-class models, weighted by the posterior probability of class-membership. We estimate the model with $h \in \{1, \ldots, K\}$ latent classes and report the best performance measure from these $K = 10$ models for every performance metric that we introduce in the upcoming subsection.

The second benchmark model that also captures heterogeneity in customer preferences is the random parameters logit (RPL) choice model (see Appendix A2 for the details), that assumes that each customer in each purchase instance samples the $\beta$ parameters of the product utilities according to some distribution and then make a choice according to a single-class MNL model with parameter vector $\beta$. In comparison with LC-MNL benchmark, RPL model allows us to assume that parameter vector $\beta$ is continuous. In particular, we assume that parameter vector $\beta$ is sampled according to multivariate normal distribution with mean $\mu$ and diagonal variance-covariance matrix $\Sigma$, i.e., $\beta \sim N(\mu, \Sigma)$. Calibration of the RPL choice model is based on the sample average approximation approach, which is computationally intensive.
4.3. Prediction Performance Measures

In general, we want to predict the product purchased by customer \( i \) in period time \( t + 1 \) given the sales transaction data of the customer up to period \( t \) and the set of items the customer is offered at time \( t + 1 \). In particular, we compare the models based on a one-step ahead prediction experiment for every category under two different metrics: “chi-square” (i.e., \( \chi^2 \)) and miss rate. For each category of products under PO-MNL Promotion model we have the following three subsets of customers: customers without cycles in their preference graph after Phase 1, customers with cycles in their preference graph, and combination of all the customers. Then for every specified subset of customers we estimate the PO-MNL Promotion model against LC-MNL and RPL benchmark models.

First, consider “chi-square” score metric, computed as follows:

\[
\chi^2 \text{ score} = \frac{1}{|\mathcal{N}||U|} \sum_{i \in \mathcal{U}, j \in \mathcal{N}} \frac{(n_{ij} - \hat{n}_{ij})^2}{0.5 + \hat{n}_{ij}}, \text{ where } \hat{n}_{ij} = \sum_{t=1}^{T_i} f_i(j, t),
\]

where \( U \) is the set of all individuals, \( n_{ij} \) is the observed number of times individual \( i \) purchased product \( j \) during the time horizon of length \( T \), \( f_i(j, t) \) takes value 1 if product \( j \) has the highest choice probability for individual \( i \) at time \( t \), and 0 otherwise. This score metric measures the ability of the different models to predict the aggregate market shares for every individual, where lower scores indicate better predictive accuracy.

Second, we compare the different models based on the miss rate computed as

\[
\text{miss rate} = \frac{1}{|U|} \sum_{i \in \mathcal{U}} \frac{1}{|T_i|} \sum_{t=1}^{T_i} \mathbb{I}[f_i(j_{it}, t) = 1],
\]

where \( f_i(j_{it}, t) \) takes value 1 if product \( j_{it} \) has the highest choice probability for individual \( i \) at time \( t \), and 0 otherwise; \( \mathbb{I}[A] \) is the indicator function that takes value 1 if \( A \) is true and 0 otherwise, and \( j_{it} \) is an item purchased at time \( t \) by individual \( i \). Miss rate measures the ability of a model specification to predict the purchase for every transaction separately. Note that both scores are reflective of the practical prediction problem that every retailer faces.

4.4. Brand Choice Prediction Results

We compare the predictive performance of the PO-MNL Promotion model with two benchmarks (LC-MNL and RPL), based on “chi-square” score and miss rate. Note that the product promotion information is taken into account in all the compared models. In addition to PO-MNL Promotion model (single class model), we estimate the LC PO-MNL Promotion model (multi-class model) with \( h \in \{1, \ldots, K\} \) latent classes and report the best performance measure from these \( K = 10 \) models for both performance metrics.
Figure 2 presents scatterplots of the “chi-square” scores of LC-MNL and RPL versus “chi-square” scores of PO-MNL Promotion (single class) and LC PO-MNL Promotion (multi-class), across the 27 product categories, for three subsets of customers. We conclude that PO Promotion method outperforms both LC-MNL and RPL benchmark models across the board. First, we consider the left column in Figure 2, where we calibrate the models for the subset of individuals that do not have cycles in their preference graph, comparing PO-MNL Promotion and LC PO-MNL Promotion models with both benchmarks. The “chi-square” score of PO-MNL Promotion model exhibits an average improvement of 10.25% over LC-MNL and 4.55% over RPL. This improvement in prediction performance can be explained by the effectiveness of the DAGs to capture partial preferences of the customers. It is an influential result since both benchmark models have more parameters to estimate and require around $300 \times$ more time it takes to estimate PO-MNL Promotion model. The key attribute of PO-MNL Promotion model making it superior to the benchmarks is that it accounts for heterogeneous customer preferences through their partial orders, so that it makes more efficient use of the limited purchase transaction data. Considering LC PO-MNL Promotion model, we can further boost the performance having an average improvement of 14.72% over LC-MNL and 8.98% over RPL. Second, we consider the middle column in Figure 2, where we calibrate the models for the subset of individuals that have cycles in their preference graph after Phase 1, and we obtain that PO-MNL Promotion model exhibits an average improvement of 13.01% over LC-MNL and 3.58% over RPL, while LC PO-MNL Promotion model, capturing heterogeneity of customers to a greater extent, has an average improvement of 13.96% over LC-MNL and 4.52% over RPL. Then, we consider the right column in Figure 2, that covers all the individuals while aggregating the results in the first two columns.

Figure 3 presents scatterplots of the miss rates where a display format is similar to Figure 2. It is straightforward that miss rate is a more stringent predictive measure than “chi-square” score, because it is based on every individual transaction assessment. In the Figure 3 we observe that our model combinations obtain improvements of between 0.05% and 4.01% under PO-MNL Promotion, and further improvements of between 2.36% and 6.48% under LC PO-MNL Promotion over the benchmarks in all six panels.

To conclude, we have several observations. First, the preference decycling framework of accounting for unobservable consideration sets further boosts the improvement of PO-based models over the classical benchmarks by increasing the coverage of individuals to 100%. Second, from all the panels it can be concluded that RPL model outperforms the LC MNL model on average across 27 categories of products. Third, for all the panels we have that LC PO-MNL model boosts the performance of PO-MNL model by accounting for additional heterogeneity of customers. Last, we have that PO-MNL Promotion model (or LC PO-MNL model) outperforms all the other model
Figure 2 Scatter plot of the average $\chi^2$ scores of all 27 product categories under the PO-MNL Promotion [single class] and LC PO-MNL Promotion [multiclass], vs. the best of up to 10 LC-MNL [benchmark] and RPL [benchmark]. Lower is better; the benchmark is outperformed for points above the 45° line. Left panels: we consider only individuals that do not have cycles in the preference graph under the Promotion model. Middle panels: we consider only individuals that have cycles in the preference graph under the Promotion model. Right panels: combine both types of individuals.

combinations including LC MNL and RPL benchmarks that incorporate the same information of promotions. Therefore, this model can be used to measure customer response to product promotions when we have very few observations for each customer by capturing partial preferences of the customers by DAGs. In general, PO-based models outperform the benchmarks especially when we have the problem of data sparcity.

5. Optimization of Personalized Promotions

We consider promotions from the retailer’s standpoint. In this case each time a customer visits the store, the retailer chooses the optimal set of products to promote. By modeling the choice behavior of a customer, we could predict which product would be purchased with the highest probability. We formulate the optimization problem to determine which items to promote under PO-MNL Promotion model.

The optimization problem formulated below has to be solved each time the customer visits the store. Recall that under PO-MNL Promotion model each product $j \in \{1, 2, ..., n\}$ is represented
by its nonpromoted copy $a_j$ and promoted copy $a_{j+n}$. For all $j \in \{1, 2, ..., n\}$ we define $r_i$ as the revenue of nonpromoted product $a_j$, and $d_j$ as the absolute value of the price discount when the product is under promotion. For the no purchase option we have that $r_0 = 0$. Let $C$ denote the available set of items where every item is represented by two copies: promoted and nonpromoted copies, and $D$ is the DAG of a customer. Then we further introduce some notation: $R_j = r_j$, $v_j = \exp(\beta_j)$ for all $j \in \{0, 1, ..., n\}$, $R_j = r_{j-n} - d_{j-n}$, $v_j = \exp(\beta^0_j + \beta_j)$ for all $j \in \{n + 1, ..., 2n\}$. Note that we complete transitive closures in DAG $D$ with the Floyd-Warshall algorithm before running personalized promotion optimization. In particular, we add an edge between products $a_j$ and $a_k$ whenever there is a directed path between these items in the original DAG $D$. The transitive closure of any DAG with $O(n)$ nodes can be computed within $O(n^3)$ computational time. Next, we formulate the personalized promotion algorithm as a MILP.
5.1. Promotion Optimization: MLIP Formulation

For each store visit, we need to decide which items that are available in the store to put under promotion. For all $j \in \{1, \ldots, n\}$ s.t. $a_j \in C$ we have that

$$y_j + y_j + n = 1 - y_j = \begin{cases} 1, & \text{if item } j \text{ is put under promotion, i.e., copy } a_j + n \in C \text{ is offered,} \\ 0, & \text{otherwise, i.e., copy } a_j \in C \text{ is offered.} \end{cases}$$

Note that the vector $y$ specifies the assortment $S$ of product copies offered to a customer, i.e., binary vector $y$ characterizes the offer set $S$. Each time a customer visits the store, the retailer chooses the items to put on promotion to maximize the revenue from the product that is most likely to be purchased by a customer according to the PO-MNL Promotion model:

$$\max_y \sum_{a_j \in C: j \leq n} \left( r_j I[y_j = 1] f(a_j, y, D) + (r_j - d_j) I[y_j + n = 1] f(a_j + n, y, D) \right)$$

s.t. $y_j + y_j + n = 1, \forall a_j \in C, j \leq n,$

$y_i \in \{0, 1\}, \forall a_i \in C,$

where $I[f(a_j, y, D)]$ takes value 1 if product $a_j \in C$ has the highest choice probability, i.e., $f(a_j, y, D)$. For all $a_j \in N'$ we have that

$$f(a_j, y, D) = \begin{cases} \frac{v_{D}(a_j)}{\sum_{t \in h_D(y)} v_{D}(t)}, & \text{if } a_j \in h_D(y), \\ 0, & \text{otherwise,} \end{cases}$$

where $v_{D}(a_j) = \sum_{a_l \in \Psi_D(a_j)} v_l; h_D(y)$ are heads of the subgraph induced in $D$ by the set $S = \{a_j \in C : y_j = 1\}$. Next we formulate an equivalent integer program to choose the optimal set of items to promote under the PO-MNL Promotion choice model. First, we define

$$B_{kj} = \begin{cases} 1, & \text{if } (a_k, a_j) \in E_D \\ 0, & \text{otherwise,} \end{cases}$$

Also we define auxiliary decision variables $z_j$ for the promotion optimization problem to specify the set of heads of the subgraph induced in $D$ by the set of offered items $S$, i.e., $h_D(y)$:

$$z_j = \begin{cases} 1, & \text{if item } a_j \in h_D(y), \\ 0, & \text{otherwise.} \end{cases}$$

Then it is straightforward to reformulate the promotion optimization problem as follows:

$$\max_{y, z} \sum_{a_k \in C} R_k f_k(z), \text{ where } \begin{cases} f_k(z) = \begin{cases} 1, & \text{if } \arg \max_{a_p \in C} \frac{v_{D}(a_p) z_p}{\sum_{a_j \in C} v_{D}(a_j) z_j} = a_k \\ 0, & \text{otherwise,} \end{cases} \end{cases}$$
\[ z_j = \begin{cases} 1, & \text{if } \sum_{a_k \in C \setminus \{a_j\}} B_{kj} r_k = 0 \text{ and } y_j = 1, \\ 0, & \text{otherwise}, \end{cases} \quad (C1) \]

\[ y_j + y_{j+n} = 1, \quad \forall a_j \in C, \quad j \leq n, \quad (C2) \]

\[ z_j, y_j \in \{0,1\}, \quad \forall a_j \in C. \quad (C3) \]

Next, we formulate a more tractable promotion optimization problem below, which is equivalent to the optimization problem (7) according to Lemma A4:

\[
\max_{y,z} \sum_{a_k \in C} R_k f_k(z), \quad (8)
\]

\[
f_k(z) = \begin{cases} 1, & \text{if } \arg \max_{a_p \in C} \frac{v_{\Psi D(a_p)} z_p}{\sum_{a_j \in C} \Psi D(a_j) z_j} = a_k \\ 0, & \text{otherwise}, \end{cases}
\]

s.t. \[ z_j \leq 1 - B_{kj} y_k, \quad \forall a_k, a_j \in C, \quad k \neq j, \quad (C4) \]

\[ y_j \leq \sum_{a_k \in C \setminus \{a_j\}} B_{kj} r_k + z_j, \quad \forall a_j \in C, \quad (C5) \]

\[ z_j \leq y_j, \quad \forall a_j \in C, \quad (C6) \]

\[ y_j + y_{j+n} = 1, \quad \forall a_j \in C, \quad j \leq n, \quad (C7) \]

\[ z_j, y_j \in \{0,1\}, \quad \forall a_j \in C. \quad (C8) \]

The first set of constraints (C4) in the optimization problem above imply that if item \( a_j \) belongs to the set \( h_D(y) \) then it does not have an incoming edge in the DAG \( D \) from other offered items. The second set of constraints (C5) ensure that every offered item \( a_j \in C \) is either belongs to the set \( h_D(y) \) or has an incoming edge in the DAG \( D \) from some other offered items. According to the third set of constraints (C6), if \( a_j \in C \) is not offered to a customer, then it can not belong to the set \( h_D(y) \). The last set of equalities (C7) say that all the products that are available in the store are offered either under full price or promoted price. The integer optimization problem above is not a MILP, however, based on the next lemma, we can formulate it as a MILP in the following way:

\[
\max_{y,z,p,\alpha} \sum_{a_j \in C} R_j \alpha_j \quad (9)
\]

s.t. \[ z_j \leq 1 - B_{kj} y_k, \quad \forall a_k, a_j \in C, \quad k \neq j, \quad (C9) \]

\[ y_j \leq \sum_{a_k \in C \setminus \{a_j\}} B_{kj} y_k + z_j, \quad \forall a_j \in C, \quad (C10) \]

\[ z_j \leq y_j, \quad \forall a_j \in C, \quad (C11) \]

\[ y_j + y_{j+n} = 1, \quad \forall a_j \in C, j \leq n, \quad (C12) \]

\[ 0 \leq p_j \leq v_{\Psi D(a_j)} p_0, \quad \forall a_j \in C, \quad (C13) \]
\[ \sum_{a_j \in C} p_j = 1, \quad \text{(C14)} \]

\[ p_j \leq z_j, \quad \forall a_j \in C, \quad \text{(C15)} \]

\[ v_{\Psi_D(a_j)} p_0 + v_{\Psi_D(a_j)} z_j - v_{\Psi_D(a_j)} \leq p_j, \quad \forall a_j \in C, \quad \text{(C16)} \]

\[ \alpha_k \leq 1 + p_k - p_j, \quad \forall a_k, a_j \in C, k \neq j, \quad \text{(C17)} \]

\[ \sum_{a_j \in C} \alpha_j = 1, \quad \text{(C18)} \]

\[ y_j, z_j, \alpha_j \in \{0, 1\}, \quad \forall a_j \in C. \quad \text{(C19)} \]

**Lemma 1.** Promotion optimization problems (8) and (9) have the same optimal objective value and the optimal solution of one of these problems can be used to find the optimal solution to the other.

### 5.2. Personalized Promotions: Performance Evaluation

It is assumed that price is an exogenous variable in our model and the promotion decision is binary, i.e., the full price of an item and the discounted price of an item are inputs for our model. We calculate the full price (or discounted price) of each item as an average price (or an average discounted price) of the item over the training data.

We test the algorithm to run customized promotions on the real data described in Section 4.1. For each store visit we are given the set of items offered to a customer, which consists of items under promotion and under full price. Having this information, we firstly calculate the expected revenue per store visit under PO-MNL Promotion model when we do not optimize promotions. Next, we obtain the set of items to promote by solving the promotion optimization problem (9) and using this optimal set of items to promote we compute the expected revenue of the retailer per customer’s visit under PO-MNL Promotion choice model. From the results shown in the left panel of Figure 4, we conclude that running personalized promotions the retailer can increase the revenue by 26.48% on average across 27 categories of products based on PO-MNL Promotion model. In the right panel of Figure 4, we report the loyalty score of 27 categories of products. We compute the loyalty score of every product category by taking the average percentage of times a customer buys her most frequently purchased product (i.e., vendor) from the category. In the middle panel of Figure 4, we also investigate the correlation between the percentage improvement of retailer’s revenue after promotion optimization based on PO-MNL Promotion versus the loyalty scores. There we could see a clear negative correlation between the percentage revenue improvement and loyalty scores. Therefore, we conclude that the retailer has the highest potential to extract the revenue from categories of products with lower loyalty score.
5.3. Inferences from PO-MNL Promotion Model

We show that PO-MNL Promotion model has a better predictive performance than classical benchmark models, so that it can be used to measure a customer response to product promotions. We calibrate the model and then run the personalized promotion algorithm to find out which items to promote for every individual. Another good property of this model is that it is very illustrative. In particular, DAG structure gives an intuitive, tractable and systematic way to make some inferences that could help the modeler to predict a consumer’s response to promotions. Note that even without fitting the model and without running the personalized promotion algorithm, we can already make some DAG-based observations of which products to promote for each individual. For example, assume that a customer $i$ has a DAG $D$ depicted in the Figure 1. From this figure alone we could infer several observations that could be of great importance for the retailer while offering personalized deals to the individual $i$: (1) we need to put item 4 under promotion for individual $i$ in order to induce its purchase; (2) it is not worth promoting item 3 for individual $i$, since it is never going to be purchased; (3) we need to put item 1 under promotion for individual $i$ in order to induce its purchase if item 2 is not under promotion; (4) it is not worth promoting item 1 for individual $i$ if item 2 is under promotion, since it is never going to be purchased in this case; etc.
6. Conclusion

Sales promotions planning is an important part of day-to-day operations in the retail industry, where the huge proportion of products is sold under discounted prices. For many years grocery retailers have been running mostly massive promotions, offering the same deal to all the customers, due to its simple practical implementation. Consequently, the negative effect of price promotions (e.g., stockpiling) might dominate the outcome of massive promotions. Since different customers are effected by promotions differently, it is worth for the retailer to offer personalized deals, which become practical nowadays because of the unprecedented volume of panel data on the sales transactions that businesses are able to collect and new technology to personalize the customer experience. As a result, personalization can mitigate the negative effects of promotions and be used as a perfect tool towards price discrimination.

In the paper we consider a back-to-back methodology to run personalized promotions to increase the retailers revenue by inducing the brand switching effect. Note that sales of one product can be reduced when another product is on promotion as some customers might switch from one brand to another. Thus, the important step in personalized promotion planning is to develop an accurate choice-based demand model and then use it to measure the customer response to product promotions.

In order to quantify the promotion response of customers, we analyze a nonparametric partial order based demand estimation model that explicitly accounts for promotions by creating two copies for every item in the product category: promoted and nonpromoted versions. Every customer is described by a set of partial preferences which is the collection of pairwise comparisons of various product copies. Consequently, the choice process proceeds as follows: (1) upon arrival to the store, a customer samples a full preference list of all the product copies in the category consistent with her partial order; (2) the customer purchases the top rank item from the considered set of products. Since inference of customers’ partial orders from sales transactions data depends on the consideration set definition, we propose a novel approach to estimate partial orders, efficiently modeling consideration set formation in a data-driven way. From an empirical perspective, we calibrate the proposed nonparametric choice-based demand model and demonstrate its ability to make more precise and fine-grained predictions of customers’ responses to price promotions compared to state-of-the-art benchmarks. Theoretically, we derive tractable lower and upper bounds relative to the exact likelihood of partial orders (which is hard to compute). These bounds are able to provide the estimation and prediction guarantees of the class of partial order based choice models. Then, using the proposed choice modeling framework we formulate a personalized promotion optimization algorithm from the retailer’s perspective to induce a brand switching behavior of customers. Our results provide a solid basis for the practical implementation of the proposal in the retail sector.
Endnotes

1. According to Bloomberg, Unata Inc. provides an analytics and marketing platform that connects loyalty/CRM platform of the retailer to digital channels to enable a personalized and real-time customer experience.

References


Customized individual promotions: Model, optimization, and prediction

APPENDIX

Srikanth Jagabathula
NYU Stern School of Business, New York, sjagabat@stern.nyu.edu

Dmitry Mitrofanov
NYU Stern School of Business, New York, dm3537@stern.nyu.edu

Gustavo Vulcano
NYU Stern School of Business, New York, gvulcano@stern.nyu.edu
School of Business, Universidad Torcuato di Tella, Buenos Aires, Argentina

Notation. We now summarize the relevant notation from the main body and also introduce additional notation for this section. Let $N = \{a_1, \ldots, a_n\}$ denote the universe of $n$ products. For the purposes of this section, we ignore the no purchase option and information about product promotions in order to simplify notation and exposition. This assumption is without loss of generality since we can account for promotions and include the no purchase option by expanding the product universe.

A DAG $D \subseteq \{(a_j, a_{j'}) : 1 \leq j, j' \leq n\}$ is a collection of pairwise preferences. We visualize a DAG $D$ as a directed graph with nodes as products and a directed edge from $a$ to $b$ if $(a, b) \in D$. We abuse notation and let $D$ denote both the directed graph and the collection of pairwise preferences. We let $E_D$ denote the set of pairwise preferences in the transitive reduction of $D$.

Let $S_n$ denote the collection of all possible $n!$ rankings or permutations of the products in $N$. For any ranking $\sigma \in S_n$, $\sigma(a)$ denotes the preference ranking of product $a$ under ranking $\sigma$ such that product $a_j$ is preferred over product $a_{j'}$ under $\sigma$ if and only if $\sigma(a_j) < \sigma(a_{j'})$. Given a DAG $D$, let $S_D \subseteq S_n$ denote the set of rankings that are consistent with $D$, i.e., $S_D = \{\sigma \in S_n : \sigma(a) < \sigma(b) \text{ whenever } (a, b) \in D\}$.

For any product $a_j$ and DAG $D$, the reachability set $\Psi_D(a)$ comprises the set of all nodes that can be reached from $a$ through a directed path in $D$, i.e., $\Psi_D(a) := \{b :$ there is a directed path from $a$ to $b$ in $D\}$. We assume that $a$ is reachable from itself, so $a \in \Psi(a)$. On the other hand, $\Theta_D(a)$ comprises the set of nodes from which $a$ can be reached, i.e., $\Theta_D(a) := \{b :$ there is a directed path from $b$ to $a$ in $D\}$. To be consistent, we also include $a$ in $\Theta_D(a)$. When the DAG $D$ is clear from the context, we drop $D$ from the notation and simply write $\Psi(a)$ and $\Theta(a)$. 
For any subset $S \subseteq \mathcal{N}$, suppose $\pi$ is a ranking of the products in $S$. Then, $\sigma(\pi)$ denotes the set of all complete rankings of the products in $\mathcal{N}$ that are consistent with $\pi$, i.e., $\sigma(\pi) = \{\sigma \in \mathcal{S}_n : \sigma(a) < \sigma(b) \text{ whenever } \pi(a) < \pi(b)\}$.

We now establish the following results.

**Lemma A1.** Given a DAG $D$, let $a_y \in \mathcal{N}$ be a node such that every node in $\Psi_D(a_y) \setminus \{a_y\}$ has at most one incoming edge and the subgraph $D[a_y]$ induced in $D$ by the set of nodes $\Psi_D(a_y)$ is a directed tree. Further, let $\bar{D}[a_y]$ denote the subgraph induced in $D$ by the set of nodes $(\mathcal{N} \setminus \Psi_D(a_y)) \cup \{a_y\}$. Then, under the MNL distribution $\lambda$, we have that

$$\lambda(D) = \lambda(D[a_y]) \cdot \lambda_y(\bar{D}[a_y]),$$

where $\lambda_y$ is the distribution over rankings obtained by replacing the MNL weight $v_y$ of product $a_y$ with $v_{\Psi_D(a_y)}$.

**Proof of Lemma A1:** Note that $D[a_y]$ is the tree “hanging” from the node $a_y$ in DAG $D$. We establish the result of this lemma by showing that the term $\lambda(D[a_y])$ factors out from the expression for $\lambda(D)$.

For that, let $S_1$ denote $\mathcal{N} \setminus \Psi_D(a_y)$ and $S_2$ denote $\Psi_D(a_y)$. It is clear that $S_1 \cap S_2 = \emptyset$ and $S_1 \cup S_2 = \mathcal{N}$. For any ranking $\sigma$ and position $1 \leq r \leq n$, let $\sigma^{-1}(r)$ denote the product ranked at position $r$ under $\sigma$. Let $D_1$ and $D_2$ denote the subgraphs in $D$ induced by the sets $S_1$ and $S_2$, respectively. It follows from our notation that $D_2 = D[a_y]$. Let $E_1$ and $E_2$ denote the edges in the transitive reductions of $D_1$ and $D_2$, respectively.

With this notation, we first claim that the set $E_D$ of edges in the transitive reduction of $D$ can be partitioned as

$$E_D = E_1 \cup E_2 \cup E_3, \text{ where } E_3 = \{(a, a_y) : (a, a_y) \in E_D\} \text{ and } E_i \cap E_j = \emptyset \forall 1 \leq i \neq j \leq 3 \quad (A1)$$

To establish this claim, we first note that $E_1 \cup E_2 \cup E_3 \subseteq E_D$ since it follows by definition that $E_i \subseteq E_D$ for all $1 \leq i \leq 3$. To show that $E_D \subseteq E_1 \cup E_2 \cup E_3$, consider an edge $(a, b) \in E_D$. We note that if $a \in S_2$, then $b$ must belong to $S_2$. The reason is that since $S_2 = \Psi_D(a_y)$, if $a \in S_2$, then $a$ is reachable from $a_y$ and because $b$ is reachable from $a$, it must be that $b$ is also reachable from $a_y$, which implies that $b \in \Psi_D(a_y) = S_2$. Therefore, there are two cases to consider: (i) both $a$ and $b$ belong to $S_1$ or $S_2$ and (ii) $a \in S_1$ and $b \in S_2$. In case (i), it follows by definition that $(a, b)$ belongs to $E_1$ or $E_2$. In case (ii), since every node in $S_2 \setminus \{a_y\}$ can have at most one incoming edge and every node in $S_2 \setminus \{a_y\}$ already has an incoming edge from a node in $S_2$, the only way there can be an edge from $a \in S_1$ to $b$ is if $b = a_y$. It now follows that $(a, b) = (a, a_y) \in E_3$. We have thus shown
that \( E_D = E_1 \cup E_2 \cup E_3 \). The fact that the three sets \( E_1, E_2, \) and \( E_3 \) are mutually disjoint follows immediately from noting that the sets \( S_1 \) and \( S_2 \) are disjoint.

With the above decomposition of the edges of \( E_D \), we now show that the set of rankings \( S_D \) can be decomposed in a convenient manner. For that, let \( \pi_1 \) and \( \pi_2 \) be rankings of products in the sets \( S_1 \) and \( S_2 \), respectively. Let \( X \) be the set of tuples \((\pi_1, \pi_2)\) such that \( \pi_1 \) and \( \pi_2 \) are consistent with DAGs \( D_1 \) and \( D_2 \), respectively. In other words, \( X = \{(\pi_1, \pi_2) : \sigma(\pi_1) \subseteq S_{D_1}, \sigma(\pi_2) \subseteq S_{D_2}\} \).

For any \( 1 \leq i \leq k_1 \), let \( S_i(\pi_1, \pi_2) \) denote the set of rankings in \( S_i \) consistent with both \( \pi_1 \) and \( \pi_2 \) where the head of \( \pi_2 \) is located after the \( i \)-th element of \( \pi_1 \), i.e., \( S_i(\pi_1, \pi_2) = \{ \sigma \in S_i : \sigma \in \sigma(\pi_1) \cap \sigma(\pi_2), \text{ with } \sigma(\pi_2^{-1}(1)) \geq \sigma(\pi_1^{-1}(i)) + 1 \} \). Further, let \( i(\pi_1) \) denote the position of the least preferred item in \( \Theta_D(a_y) \) in the ranking \( \pi_1 \), i.e., \( i(\pi_1) = \max \{ \pi_1(a) : a \in \Theta_D(a_y) \} \). We claim that the set of rankings \( S_D \) is obtained by taking a tuple \((\pi_1, \pi_2) \in X \) and combining them such that the head of \( \pi_2 \) occurs after the \( i(\pi_1) \)-th element of \( \pi_1 \). More precisely, we claim that

\[
S_D = \bigcup_{(\pi_1, \pi_2) \in X} S_i(\pi_1, \pi_2) \text{ where } S_i(\pi_1, \pi_2) \cap S_i(\pi_1', \pi_2') \text{ for } (\pi_1, \pi_2) \neq (\pi_1', \pi_2')
\]  

(A2)

To establish the claim, we first show that \( S_i(\pi_1, \pi_2) \subseteq S_D \) for all \((\pi_1, \pi_2) \in X\). For that, consider a \( \sigma \in S_i(\pi_1, \pi_2) \) and consider an edge \((a, b) \in E_D\). It is sufficient to show that \( \sigma(a) < \sigma(b) \). It follows from (A1) that \((a, b)\) is in either \( E_1 \) or \( E_2 \) or \( E_3 \). If \((a, b)\) is in \( E_1 \), then we must have that \( \pi_1(a) < \pi_1(b) \) because \( \pi_1 \) is consistent with \( D_1 \). Since \( \sigma \) is consistent with \( \pi_1 \), we have that \( \sigma(a) < \sigma(b) \). Using a symmetric argument, we can similarly show that \( \sigma(a) < \sigma(b) \) when \((a, b) \in E_2 \).

Now suppose that \((a, b) \in E_3 \). We then have that \( b = a_y \). Since \( \pi_2 \) is consistent with \( D_2 \) and \( a_y \) is preferred over every product in \( S_2 \setminus \{a_y\} \) under the partial order \( D_2 \), it follows that \( a_y \) must be the head of \( \pi_2 \), i.e., \( \pi_2(a_y) = 1 \). Now, let \( a^* \) denote the least preferred element under \( \pi_1 \) in the set \( \Theta_D(a_y) \). Since \((a, a_y) \in E_D \), we have that \( a \in \Theta_D(a_y) \), implying that \( \sigma(a) = \pi_1(a) < \pi_1(a^*) = \sigma(a^*) \). It also follows by our definitions that \( \pi_1(a^*) = i(\pi_1) \) and \( \sigma(a_y) = \sigma(\pi_2^{-1}(1)) > \sigma(\pi_1^{-1}(i(\pi_1))) = \sigma(a^*) \), where the inequality follows from the definition of \( S_i(\pi_1, \pi_2) \) and the fact that \( \sigma \in S_i(\pi_1, \pi_2) \). We have thus shown that \( \sigma(a) < \sigma(a^*) < \sigma(a_y) = \sigma(b) \), establishing the result that \( S_i(\pi_1, \pi_2) \subseteq S_D \) for all \((\pi_1, \pi_2) \in X \), which implies that \( \bigcup_{(\pi_1, \pi_2) \in X} S_i(\pi_1, \pi_2) \subseteq S_D \).

We now show that \( S_D \subseteq \bigcup_{(\pi_1, \pi_2) \in X} S_i(\pi_1, \pi_2) \). For that consider a \( \sigma \in S_D \) and let \( \pi_1 \) and \( \pi_2 \) denote the rankings \( \sigma \) induces on the set of products \( S_1 \) and \( S_2 \), respectively. We show that \( \sigma \in S_i(\pi_1, \pi_2) \). It follows by the definitions of \( \pi_1 \) and \( \pi_2 \) that \( \sigma \in \sigma(\pi_1) \cap \sigma(\pi_2) \). Therefore, it is sufficient to show that \( \sigma(\pi_2^{-1}(1)) \geq \sigma(\pi_1^{-1}(i(\pi_1))) + 1 \). Using the arguments above, it readily follows that \( \pi_2^{-1}(1) = a_y \). Since \( \sigma \) is consistent with \( D \), we have that \( \sigma(a) < \sigma(a^*) \) for all \( a \in \Theta_D(a_y) \), and in particular, \( \sigma(a^*) < \sigma(a_y) \), where \( a^* \) is the least preferred product under \( \pi_1 \) from the set \( \Theta_D(a_y) \). Since \( i(\pi_1) = \pi_1(a^*) \) by definition, we have shown that \( \sigma(\pi_1^{-1}(i(\pi_1))) < \sigma(a_y) \). We have thus established that \( S_D \subseteq \bigcup_{(\pi_1, \pi_2) \in X} S_i(\pi_1, \pi_2) \).
The arguments above establish that \( S_D = \bigcup_{(\pi_1, \pi_2) \in X} S_{i(\pi_1)}(\pi_1, \pi_2) \). The disjointness of \( S_{i(\pi_1)}(\pi_1, \pi_2) \) and \( S_{i(\pi'_1)}(\pi'_1, \pi'_2) \) for \((\pi_1, \pi_2) \neq (\pi'_1, \pi'_2)\) readily follows from the disjointness of \( \sigma(\pi_1) \cap \sigma(\pi_2) \) and \( \sigma(\pi'_1) \cap \sigma(\pi'_2) \) for \((\pi_1, \pi_2) \neq (\pi'_1, \pi'_2)\). We have thus established the claim in (A2).

We can now write from (A2) that

\[
\lambda(D) = \sum_{\sigma \in S_D} \lambda(\sigma) = \sum_{(\pi_1, \pi_2) \in X} \lambda(S_{i(\pi_1)}(\pi_1, \pi_2)) = \sum_{\pi_1: \sigma(\pi_1) \subseteq S_{D_1}} \sum_{\pi_2: \sigma(\pi_2) \subseteq S_{D_2}} \lambda(S_{i(\pi_1)}(\pi_1, \pi_2)).
\]

Now consider a \((\pi_1, \pi_2) \in X\). Without loss of generality, suppose that \( \pi_1 = (a_1, 1, \ldots, a_{k_1}, 1) \) and \( \pi_2 = (a_2, 1, \ldots, a_{k_2}, 1) \). Then, invoking (Jagabathula and Vulcano 2017, Lemma A1), we can write

\[
\lambda(S_{i(\pi_1)}(\pi_1, \pi_2)) = \left[ \prod_{j=1}^{v(\pi_1)} \sum_{j'=j}^{k_1} v_{1,j} + \sum_{a_2, j' \in S_2} v_{2,j'} \right] \cdot \lambda(\pi_2) = g(\pi_1) \cdot \lambda(\pi_2),
\]

where the second equality follows from the fact that \( S_2 = \Psi_D(a_y) \) and \( g(\pi_1) \) is defined to be the term in the square brackets. We now have

\[
\lambda(D) = \sum_{\pi_1: \sigma(\pi_1) \subseteq S_{D_1}} \sum_{\pi_2: \sigma(\pi_2) \subseteq S_{D_2}} \lambda(S_{i(\pi_1)}(\pi_1, \pi_2)) = \left[ \sum_{\pi_1: \sigma(\pi_1) \subseteq S_{D_1}} g(\pi_1) \right] \cdot \left[ \sum_{\pi_2: \sigma(\pi_2) \subseteq S_{D_2}} \lambda(\pi_2) \right].
\]

Noting that

\[
\sum_{\pi_2: \sigma(\pi_2) \subseteq S_{D_2}} \lambda(\pi_2) = \sum_{\sigma \in S_{D_2}} = \lambda(D_2) = \lambda(D[a_y]),
\]

we have shown that

\[
\lambda(D) = \lambda(D[a_y]) \cdot \left[ \sum_{\pi_1: \sigma(\pi_1) \subseteq S_{D_1}} g(\pi_1) \right]. \tag{A3}
\]

It now suffices to show that

\[
\lambda_y(D[a_y]) = \sum_{\pi_1: \sigma(\pi_1) \subseteq S_{D_1}} g(\pi_1).
\]

For that, consider the distribution \( \lambda_y \) in which the weight \( v_y \) is replaced by \( v_{\Psi_D(a_y)} \), and repeat the above set of arguments for the DAG \( \overline{D}[a_y] \) with the nodes of the DAG decomposed into sets \( S_i \) as defined above and \( S_3 = \{a_y\} \). For any ranking \( \pi_1 \) of the products in set \( S_1 \), note that \( g(\pi_1) \) remains the same under both distributions \( \lambda \) and \( \lambda_y \). As a result, following (A3), we can write

\[
\lambda_y(D[a_y]) = \lambda_y(D_3) \cdot \left[ \sum_{\pi_1: \sigma(\pi_1) \subseteq S_{D_1}} g(\pi_1) \right],
\]
where $D_3$ is the DAG induced in $D$ by $S_3$. Since $S_3$ is a singleton, the DAG $D_3$ will be empty, implying that $\lambda_y(D_3) = 1$. We have thus shown that

$$\lambda_y(D[y]) = \sum_{\pi_1: \sigma(\pi_1) \subset S_D} g(\pi_1).$$

The result of the lemma now follows.

To establish the next result, we need the following notation. Given a DAG $D$ and product $a_y$, let $\lambda^\text{aug}_y$ denote the distribution of rankings on the expanded product universe $N \cup \{a'_y\}$, where $a'_y$ is a copy of the product $a_y$, with the weight $v_y$ of product $a_y$ replaced with $v_{\Psi_D(a_y)}$ and the copy $a'_y$ assigned the weight $v_{\Psi_D(a_y)}$. We say that a node $a$ in DAG $D$ is a $v$-node if it has more than one incoming edge. We define the $v$-degree of a DAG $D$ as $\sum_{a_j \text{ is a } v\text{-node}} (d_{\text{in}} - 1)$, where $d_{\text{in}}$ is the in-degree of node $a_j$. We now establish the following result.

**Lemma A2.** Suppose that a leaf node $a_y$ in the DAG $D$ has at least two incoming edges. Then there exists DAG $D^{\text{split}}$ whose $v$-degree is one less than that of $D$, such that

$$\lambda^\text{aug}_y(D^{\text{split}}) \leq \lambda(D).$$

Furthermore the approximate likelihoods of the DAGs $D$ and $D^{\text{split}}$ are equal, i.e., $\tilde{\lambda}^\text{aug}_y(D^{\text{split}}) = \tilde{\lambda}(D)$.

**Proof of Lemma A2:** Since the leaf node $a_y$ in DAG $D$ has at least two incoming edges, suppose w.l.o.g. that $(a_1, a_y), (a_2, a_y) \in E_D$. Let $D_1$ denote the DAG obtained by adding the isolated copy $a'_y$ to $D$. Let $D^{\text{split}}$ denote the DAG obtained by erasing the edge $(a_1, a_y)$ and adding the edge $(a_2, a'_y)$ to $D_1$; in other words, $E_{D^{\text{split}}} = E_D \setminus \{(a_1, a_y)\} \cup \{(a_2, a'_y)\}$. Figure A1 illustrates these DAGs. Note that by construction, the $v$-degree of $D^{\text{split}}$ is one less than that of $D$ because the in-degree of node $a_y$ has been reduced by 1.
Then, since attraction parameters for nodes which is of the form \( \prod_{v \in N} r_v \) with the weights of the products in \( N \cup \{ a'_y \} \) remaining the same as in \( \lambda \) and the weight \( v_y \) assigned to product \( a'_y \). Now, for any ranking \( \pi \) of the products in set \( N \), let \( \sigma(\pi) \) denote the ranking of the products in the set \( N \cup \{ a'_y \} \) that is consistent with \( \pi \). By invoking (Jagabathula and Vulcano 2017, Proposition A1.2), we can show that \( \lambda(\pi) = \lambda_y^{\text{aug}}(\sigma(\pi)) \). We can now write

\[
\lambda_y^{\text{aug}}(D_1) = \sum_{\pi \in S_D} \lambda_y^{\text{aug}}(\pi) = \sum_{\pi \in S_D} \lambda_y^{\text{aug}}(\sigma(\pi)) = \sum_{\pi \in S_D} \lambda(\pi) = \lambda(D).
\]

Therefore, in order to establish that \( \lambda_y^{\text{aug}}(D^{\text{split}}) \leq \lambda(D) \), it is sufficient to show that \( \lambda_y^{\text{aug}}(D^{\text{split}}) \leq \lambda_y^{\text{aug}}(D_1) \). For that, consider the three DAGs \( I_1, I_2, \) and \( I_3 \) (see Figure A2), defined over the set of products in \( N \cup \{ a'_y \} \) such that

\[
E_{I_1} = E_{D_1} \cup \{ (a_2, a'_y) \}, E_{I_2} = E_{D_1} \cup \{ (a'_y, a_2) \}, \text{ and } E_{I_3} = (E_{D_1} \setminus \{ (a_2, a'_y) \}) \cup \{ (a_y, a_2), (a_2, a'_y) \}.
\]

It is then follows by the definitions that

\[
\lambda_y^{\text{aug}}(D_1) = \lambda_y^{\text{aug}}(I_1) + \lambda_y^{\text{aug}}(I_2) \quad \text{ and } \quad \lambda_y^{\text{aug}}(D^{\text{split}}) = \lambda_y^{\text{aug}}(I_1) + \lambda_y^{\text{aug}}(I_3).
\]

Thus, to show that \( \lambda_y^{\text{aug}}(D^{\text{split}}) \leq \lambda_y^{\text{aug}}(D_1) \), it is sufficient to show that \( \lambda_y^{\text{aug}}(I_3) \leq \lambda_y^{\text{aug}}(I_2) \).

To show that \( \lambda_y^{\text{aug}}(I_3) \leq \lambda_y^{\text{aug}}(I_2) \), define the mapping \( h : I_3 \rightarrow I_2 \) such that for any \( \sigma \in I_3 \), which is of the form \( \sigma = (\ldots, a_1, \ldots, a_y, \ldots, a_2, \ldots, a'_y, \ldots) \), we map it to \( \sigma' \) in \( I_2 \), of the form \( \sigma' = (\ldots, a_1, \ldots, a'_y, \ldots, a_2, \ldots, a_y, \ldots) \) obtained by swapping the positions of the products \( a_y \) and \( a'_y \). Now, it can be verified that the mapping \( h(\cdot) \) is an injection, i.e., \( h(\sigma) \neq h(\sigma') \) whenever \( \sigma \neq \sigma' \). Then, since attraction parameters for nodes \( a_y \) and \( a'_y \) are the same, i.e., \( \nu_{a_y} = \nu_{a'_y} \), it follows that for any \( \sigma \in I_3 \), \( \lambda_y^{\text{aug}}(\sigma) = \lambda_y^{\text{aug}}(h(\sigma)) \). As a result, we obtain

\[
\lambda_y^{\text{aug}}(I_3) = \sum_{\sigma \in I_3} \lambda_y^{\text{aug}}(\sigma) = \sum_{\sigma \in I_3} \lambda_y^{\text{aug}}(h(\sigma)) \leq \sum_{\sigma' \in I_2} \lambda_y^{\text{aug}}(\sigma') = \lambda_y^{\text{aug}}(I_2). \tag{A4}
\]
We have thus shown that $\lambda^\text{aug}_y(D^{\text{split}}) \leq \lambda^\text{aug}_y(D_1)$, which implies that $\lambda^\text{aug}_y(D^{\text{split}}) \leq \lambda(D)$.

The inequality in $\lambda^\text{aug}_y(D^{\text{split}}) \leq \lambda(D)$ is strict when all the parameters under the MNL model are positive. We establish this result by showing that there exists a $\sigma' \in I_2$ such that $h(\sigma) \neq h(\sigma')$ for all $\sigma \in I_3$. It then follows that the inequality in (A4) is strict, implying that $\lambda^\text{aug}_y(D^{\text{split}}) < \lambda(D)$.

Now, consider the ranking $\sigma' = (\ldots, a_y, \ldots, a_1, \ldots, a_2, \ldots, a_y, \ldots)$ such that $\sigma' \in I_2$. As noted above, any $\sigma \in I_3$ is of the form $\sigma = (\ldots, a_1, a_y, \ldots, a_2, \ldots, a_y, \ldots, a_y, \ldots)$, so that it gets mapped to $h(\sigma) = (\ldots, a_1, \ldots, a_y, \ldots, a_2, \ldots, a_y, \ldots)$. Therefore, we have $h(\sigma)(a_1) < h(\sigma')(a_y)$ for all $\sigma \in I_3$, whereas $\sigma'(a_1) > \sigma'(a_y)$. Thus, we have that $\sigma' \neq h(\sigma)$ for all $\sigma \in I_3$, establishing the claim.

We are now left with showing that $\tilde{\lambda}^\text{aug}_y(D^{\text{split}}) = \tilde{\lambda}(D)$. Since the reachability weights $v_{\Psi_D(a)}$, for all $a \in \mathcal{N}$, under $\lambda$, and $v_{\Psi_{D^{\text{split}}}(a)}$, for all $a \in \mathcal{N} \cup \{a_y\}$, under $\lambda^\text{aug}_y$ are equal by definition, and the approximations $\tilde{\lambda}$ and $\tilde{\lambda}^\text{aug}_y$ only depend on the reachability weights, the equality $\tilde{\lambda}^\text{aug}_y(D^{\text{split}}) = \tilde{\lambda}(D)$ immediately follows.

**Proof of Proposition 1:** We show the result, $\tilde{\lambda}(D) \leq \lambda(D)$, by induction on the $v$-degree, $k$, of DAG $D$.

**Base case:** $k = 0$. When $k = 0$, DAG $D$ does not have any $v$-nodes. It then follows from (Jagabathula and Vulcano 2017, Proposition 3.2) that $\tilde{\lambda}(D) = \lambda(D)$, establishing the base case.

**Induction hypothesis:** Suppose $\tilde{\mu}(D) \leq \mu(D)$ for any DAG $D$ with $v$-degree less than or equal to $p$, for some $p \geq 0$, for all distributions $\mu$ under the PL model.

**Induction step:** Assuming that the induction hypothesis is true, we prove the result for $k = p + 1$. It is clear that there exists a $v$-node $a_y \in \mathcal{N}$ satisfying the conditions in Lemma A1, i.e., every node in $\Psi_D(a_y) \setminus \{a_y\}$ has at most one incoming edge and the subgraph $D|a_y|$, induced in $D$ by the set of nodes $\Psi_D(a_y)$ is a directed tree. As in Lemma A1, let $\bar{D}[a_y]$ denote the subgraph induced in $D$ by the set of nodes $(\mathcal{N} \setminus \Psi_D(a_y)) \cup \{a_y\}$. Now consider

$$
\tilde{\lambda}(D) = \prod_{j \in \mathcal{N}} \frac{v_j}{\sum_{j' \in \Psi_D(a_y)} v_{j'}} 
\left( \prod_{j \in \mathcal{N} \setminus \Psi_D(a_y), a_y \in \Psi_D(a_y)} \frac{v_j}{\sum_{j' \in \Psi_D(a_y)} v_{j'}} \right) 
\left( \prod_{j \in \mathcal{N} \setminus \Psi_D(a_y), a_y \notin \Psi_D(a_y)} \frac{v_j}{v_{\Psi_D(a_y)} + \sum_{j' \in \Psi_D(a_y) \setminus \{a_y\}} v_{j'}} \right) 
= \lambda(D[a_y]) \cdot \frac{v_j}{v_{\Psi_D(a_y)} + \sum_{j' \in \Psi_D(a_y) \setminus \{a_y\}} v_{j'}}
$$

$$
= \lambda(D[a_y]) \cdot \tilde{\lambda}_y(\bar{D}[a_y]),
$$
where the fourth equation follows because $D[a_y]$ is a directed tree with a unique root, which implies that $\lambda(D[a_y]) = \tilde{\lambda}(D[a_y])$ (Jagabathula and Vulcano 2017, Proposition 3.2), and the fact that $\Psi_D(a_j) = \Psi_{D[a_y]}(a_j)$ for all $j \in \mathcal{N}$ such that $a_j \notin \Psi_D(a_j)$. We now have

$$\tilde{\lambda}(D) = \lambda(D[a_y]) \cdot \tilde{\lambda}_y(D[a_y])$$

$$\leq \lambda(D[a_y]) \cdot \tilde{\lambda}_y(D_y^{\text{split}})$$

$$\leq \lambda(D[a_y]) \cdot \lambda_y(D[a_y]), \text{ with strict inequality if } v_j > 0 \forall a_j \in \mathcal{N}$$

where the first inequality follows from induction hypothesis with distribution $\mu = \lambda_y^{\text{aug}}$ since the $v$-degree of $D_y^{\text{split}}$ is equal to $p$.

The result of the proposition now follows.

\[ \square \]

**Lemma A3.** Suppose that $\hat{D}$ is a subgraph of $D$, obtained by deleting some of the edges of $D$. Then

$$\lambda(D) \leq \lambda(\hat{D}),$$

with strict inequality under PL model if all the parameters are strictly positive.

**Proof of Lemma A3:** We must have that $S_D \subset S_{\hat{D}}$ since if $\sigma$ is consistent with $D$, i.e., $\sigma \in S_D$, then it must also be consistent with $\hat{D}$, i.e., $\sigma \in S_{\hat{D}}$. It now follows that

$$\lambda(D) = \sum_{\sigma \in S_D} \lambda(\sigma) \leq \sum_{\sigma \in S_{\hat{D}}} \lambda(\sigma) = \lambda(\hat{D}).$$

We now show that $\lambda(D) < \lambda(\hat{D})$ when all the PL parameters are strictly positive by exhibiting a ranking $\sigma \in S_D$ such that $\sigma \notin S_{\hat{D}}$. Suppose, by contradiction, we have that $S_{\hat{D}} = S_D$. Then subgraph of $\hat{D}$, which has less edges than $D$, is a transitive reduction of $D$ which results in contradiction. As a result, there is $\sigma \in S_D$ such that $\sigma \notin S_{\hat{D}}$. Then we have

$$\lambda(\hat{D}) - \lambda(D) \geq \lambda(\sigma) > 0.$$

\[ \square \]

**Proof of Proposition 3:**

$$\log R(D) = \log \left( \prod_{a \in \mathcal{N}} \frac{\sum_{a_j \in \Psi_D(a)} v_j}{\sum_{a_j \in \Psi_{\Phi(D)}(a)} v_j} \right) = \sum_{a \in \mathcal{N}} \log \left( \frac{\sum_{a_j \in \Psi_D(a)} v_j}{\sum_{a_j \in \Psi_{\Phi(D)}(a)} v_j} \right)$$

$$= \sum_{a \in \mathcal{N}} \log \left( 1 + \frac{\sum_{a_j \in \Psi_D(a) \setminus \Psi_{\Phi(D)}(a)} v_j}{\sum_{a_j \in \Psi_{\Phi(D)}(a)} v_j} \right)$$
where the first equality follows because $\Psi_D(a) = \Psi_{\Phi(D)}(a)$ for all $a$ whose reachability set $\Psi_D(a)$ has no $v$-nodes, i.e., $F_D \cap \Phi_D(a) = \emptyset$, and the last three inequalities follow from the definitions of $\Delta$, $\ell$, and $p$, respectively.

We now have that

$$0 \leq \lim_{n \to \infty} \log \frac{\lambda(D)}{\lambda(D)} \leq \lim_{n \to \infty} \log R(D) \leq \lim_{n \to \infty} p \cdot \log(1 + \ell \cdot \Delta) \leq \lim_{n \to \infty} n \cdot \log(1 + n \cdot \Delta)$$

$$= \lim_{n \to \infty} \log(1 + n \cdot \Delta) \frac{n^2 \Delta}{\Delta} = \lim_{n \to \infty} \frac{1}{\Delta} \cdot \log(1 + n \cdot \Delta) \frac{n^2 \Delta}{\Delta} = 0,$$

since $n\Delta = o(1)$ and $n^2 \Delta = o(1)$ as $n \to \infty$. \□

**Proof of Proposition 4:**

$$\log \frac{\overline{f}(a_j, S, D)}{\overline{f}(a_j, S, D)} = \log \left( \frac{\lambda(D \cup C(a_j, S))}{\lambda(D \cup C(a_j, S))} \cdot \frac{\lambda(D)}{\lambda(D)} \right)$$

$$= \log \left( \frac{\lambda(D \cup C(a_j, S))}{\lambda(D \cup C(a_j, S))} \right) + \log \left( \frac{\lambda(D)}{\lambda(D)} \right)$$

$$\leq 2 \cdot p \cdot \log(1 + \ell \cdot \Delta),$$

where the last inequality follows from Proposition 3. As argued in the proof of Proposition 3, $p \log(1 + \ell \Delta) \to 0$ as $n \to \infty$ when $\Delta n^2 = o(1)$, as $n \to \infty$. The result of the proposition now follows. \□

**A1. Optimization of Personalized Promotions**

We start this section with the lemma below where we prove the equivalence of personalized promotion optimization problems (7) and (8).

**Lemma A4.** Personalized promotion optimization problems (7) and (8) have the same optimal objective value and the optimal solution of one of these problems can be used to find the optimal solution to the other.
Proof: Since the objective function and decision variables of the optimization problems (7) and (8) are the same, it is sufficient to prove that the constraints (C4)-(C8) can be equivalently represented as the constraints (C1)-(C3), and vice versa, i.e., (C4)-(C8) ⇔ (C1)-(C3):

1. (C4)-(C8) → (C1)-(C3): Constraints (C2) and (C3) are clearly satisfied, as they also appear in (C4)-(C8). Constraint (C1) can be equivalently represented as the union of these three statements:

Statement 1: \( \sum_{a_k \in C \setminus \{a_j\}} y_k B_{kj} = 0 \& y_j = 1 \Rightarrow z_j = 1, \forall a_j \in C \). This statement is satisfied since it follows from (C5) that \( z_j = 1 \) if \( \sum_{a_k \in C \setminus \{a_j\}} y_k B_{kj} = 0 \& y_j = 1, \forall a_j \in C \).

Statement 2: \( y_j = 0 \Rightarrow z_j = 0, \forall a_j \in C \). This statement is satisfied since it follows from (C6) that \( z_j = 0 \) if \( y_j = 0, \forall a_j \in C \).

Statement 3: \( \sum_{a_k \in C \setminus \{a_j\}} y_k B_{kj} > 0 \Rightarrow z_j = 0, \forall a_j \in C \). This statement can be equivalently reformulated as the following implication: \( \exists a_k \in C \setminus \{a_j\} \) s.t. \( y_k = 1 \) and \( B_{kj} = 1 \Rightarrow z_j = 0, \forall a_j \in C \). Therefore, this statement is satisfied since it follows from (C4) that \( z_j = 0 \) if \( \exists a_k \in C \setminus \{a_j\} \) s.t. \( y_k = 1 \) and \( B_{kj} = 1, \forall a_j \in C \).

2. (C1)-(C3) → (C4)-(C8): Constraints (C7) and (C8) are clearly satisfied, as they also appear in (C1)-(C3). Then it remains to consider the constraints (C4)-(C6):

(C4): \( z_j \leq 1 - B_{kj} y_k, \forall a_k, a_j \in C, k \neq j \). For all \( a_j \in C \) these constraints are equivalently represented by the following implication: \( \exists a_k \in C, k \neq j \) s.t. \( B_{kj} y_k = 1 \Rightarrow z_j = 0 \), which, in turn, is equivalent to implication in (C1): \( \sum_{a_k \in C \setminus \{a_j\}} y_k B_{kj} > 0 \Rightarrow z_j = 0 \), i.e., (C4) follows from (C1).

(C5): \( y_j \leq \sum_{a_k \in C \setminus \{a_j\}} B_{kj} y_k + z_j, \forall a_j \in C \). For all \( a_j \in C \) these constraints are equivalently represented by the following implication: \( \sum_{a_k \in C \setminus \{a_j\}} B_{kj} y_k = 0 \& z_j = 0 \Rightarrow y_j = 0 \). Assume by contradiction that this implication does not follow from (C1)-(C3): \( \exists a_j \in C \) s.t. \( y_j = 1 \) if \( \sum_{a_k \in C \setminus \{a_j\}} B_{kj} y_k = 0 \& z_j = 0 \), which contradicts (C1).

(C6): \( z_j \leq y_j, \forall a_j \in C \). For all \( a_j \in C \) these constraints are equivalently represented by the following implication: \( y_j = 0 \Rightarrow z_j = 0 \), which follows from (C1). \( \square \)

Next we prove the Lemma 1 that personalized promotion optimization problems (8) and (9) are equivalent.

Proof of Lemma 1:

1. Optimization Problem (8) ⇒ Optimization Problem (9).

First, we prove that the optimal value of the objective function in the optimization problem (8) is feasible to the optimization problem (9). Suppose that \((y^*, z^*)\) is the fixed optimal solution to the optimization problem (8) with an objective value of \( R^* \). Then we have that the optimal retailer’s revenue is equal to the revenue of the product \( a_k^* \) that has the highest probability of being purchased under PO-MNL Promotion choice model, i.e., \( R^* = R_k^* \), where

\[
a_k^* = \arg \max_{a_m \in C} \frac{v_{ψ_D(a_m)} z_m^*}{\sum_{a_j \in C} v_{ψ_D(a_j)} z_j^*}.
\]
Then we argue that there are vectors $y', z', p', \alpha'$:

$$
\begin{align*}
    y' &= y^*, \\
    z' &= z^*, \\
    p' &= \frac{v_{\Psi D(a_k)} z^*_k}{\sum_{a_j \in C} v_{\Psi D(a_j)} z^*_j} \quad \forall a_k \in C, \\
    p' &= \frac{1}{\sum_{a_j \in C} v_{\Psi D(a_j)} z^*_j}, \\
    \alpha' &= 1 \text{ if } k = k^*, \text{ and } 0 \text{ otherwise}, \quad \forall a_k \in C,
\end{align*}
$$

such that $(y', z', p', \alpha')$ is a feasible solution to the optimization problem (9) with an objective value $R'$ such that

$$
R' = \sum_{a_k \in C} R_k \alpha' = R_{k^*} = R^*.
$$

It is straightforward to verify that $(y', z', p', \alpha')$ is a feasible solution to the optimization problem (9).

2. Optimization Problem (8) ⇔ Optimization Problem (9).

Second, we prove that the optimal value of the objective function in the optimization problem (9) is feasible to the optimization problem (8). Suppose that $(y^*, z^*, p^*, \alpha^*)$ is the fixed optimal solution to the optimization problem (9) with an objective value of $R^*$. Since the variables $z_k$ for all $a_k \in C$ are binary, we consider two cases for all $a_k \in C$:

Case 1: $z^*_k = 0$. Then it follows that $p^*_k = 0$ because of (C13) and (C15).

Case 2: $z^*_k = 1$. Then it follows that $p^*_k = v_{\Psi D(a_k)} p^*_0$ because of (C13) and (C16).

As a result, we have that $p^*_k = p^*_0 v_{\Psi D(a_k)} z^*_k$ for all $a_k \in C$, so that it follows from (C14) that

$$
p^*_0 = \frac{1}{\sum_{a_j \in C} v_{\Psi D(a_j)} z^*_j}.
$$

Then for all $a_k \in C$ we have that

$$
p^*_k = p^*_0 v_{\Psi D(a_k)} z^*_k = \frac{v_{\Psi D(a_k)} z^*_k}{\sum_{a_j \in C} v_{\Psi D(a_j)} z^*_j},
$$

Next, it follows from (C17), (C18), and (C19) that for all $a_k \in C$ we have

$$
\alpha^*_k = \begin{cases} 
1, & \text{if } k = k^*, \\
0, & \text{otherwise},
\end{cases}
$$

where $a_k^* = \arg \max_{a_m \in C} p^*_m = \arg \max_{a_m \in C} \left( \frac{v_{\Psi D(a_m)} z^*_m}{\sum_{a_j \in C} v_{\Psi D(a_j)} z^*_j} \right)$. 


Consequently, it follows that $R^* = R_{k^*}$, where $a_{k^*} = \arg\max\left(\frac{\nu_{P}(a_{m})\gamma_m}{\sum_{a_j \in C} \nu_{P}(a_j)\gamma_j}\right)$. Then we argue that there are vectors $y'$ and $z'$ such that $y' = y^*$ and $z' = z^*$, where $(y', z')$ is a feasible solution to the optimization problem (8) with an objective value $R'$ and we have that

$$R' = \sum_{a_k \in C} R_k f_k(z^*) = R_{k^*} = R^*.$$ 

It is straightforward to verify that $(y', z')$ is a feasible solution to the optimization problem (8).

\[\Box\]

### A2. Benchmark Models

#### A2.1. LC-MNL Model

LC-MNL model captures heterogeneity among customers by allowing them to belong to $K$ different classes with some probability. Customers from class $h \in \{1, ..., K\}$ make choices according to the single class MNL model with a parameter value $I_{i, t} \beta_{h_{j, i}}^0 + \beta_{h_{j, i}}$ of product $j_{i, t} \in \{1, 2, ..., n\}$, where $I_{i, t} = 1$ if product $j_{i, t}$ is under promotion at time $t$ for individual $i$, and 0 otherwise. A prior probability of a customer to belong to the class $h$ is $\gamma_h \geq 0$ such that $\sum_{h=1}^{K} \gamma_h = 1$. The regularized maximum likelihood problem under $K$ class LC-MNL model can be formulated as follows:

$$\max_{\beta, \gamma} \sum_{i=1}^{m} \log \left( \sum_{h=1}^{K} \gamma_h \prod_{t=1}^{T_i} \frac{\exp(I_{i, t} \beta_{h_{j, i}}^0 + \beta_{h_{j, i}})}{1 + \sum_{a_j \in S_{i, t}} \exp(I_{i, t} \beta_{a_j}^0 + \beta_{a_j})} \right) - \alpha \sum_{h=1}^{K} (\|\beta_h^0\|_1 + \|\beta_h\|_1)$$

When the value of $\alpha$ is fixed and $K = 1$, it can be shown that the optimization problem in (6) is globally concave and therefore can be solved efficiently (Train 2009). Note that we turned the value of $\alpha$ by 5-fold cross-validation. Since the problem is nonconcave for $K > 1$, the EM technique is used to fit the model (see Appendix A2.1.1 in Jagabathula and Vulcano 2017). Specifically, we initialize the EM with a random allocation of customers to one of the $K$ classes, resulting in an initial allocation $D_1, D_2, ..., D_K$, which form a partition of the collection of all the customers. Then we set $\gamma_h^{(0)} = |D_h|/\left(\sum_{d=1}^{K} |D_d|\right)$. In order to get a parameter vector $\tau_h^{(0)}$, we fit a single class MNL model to each subset of customers. Based on each customer $i$’s purchase history $(a_{j_{i, t}, S_{i, t}})$ for $1 \leq t \leq T_i$, we can estimate their posterior membership probabilities $\forall h \in \{1, ..., K\}$:

$$\hat{\gamma}_{ih} = \frac{\gamma_h \prod_{t \in T_i} \left[ \exp(I_{i, t} \beta_{h_{j, i}}^0 + \beta_{h_{j, i}}) / \left( 1 + \sum_{a_j \in S_{i, t}} \exp(I_{i, t} \beta_{a_j}^0 + \beta_{a_j}) \right) \right]}{\sum_{d=1}^{K} \gamma_d \prod_{t \in T_i} \left[ \exp(I_{i, t} \beta_{d_{j, i}}^0 + \beta_{d_{j, i}}) / \left( 1 + \sum_{a_j \in S_{i, t}} \exp(I_{i, t} \beta_{a_j}^0 + \beta_{a_j}) \right) \right]},$$

and the prediction can be made as follows:

$$f(j_{i, t}, S_{i, t}) = \sum_{h=1}^{K} \hat{\gamma}_{ih} \frac{\exp(I_{i, t} \beta_{h_{j, i}}^0 + \beta_{h_{j, i}})}{1 + \sum_{a_j \in S_{i, t}} \exp(I_{i, t} \beta_{a_j}^0 + \beta_{a_j})},$$

where $f(j_{i, t}, S_{i, t})$ is a probability to choose an item $j_{i, t}$ from the offer set $S_{i, t}$. 
A2.2. RPL Model

In this model, we assume that $\beta$ is sampled from multivariate normal distribution, i.e., $\beta \sim N(\mu, \Sigma)$, where $\mu$ is the mean, and $\Sigma$ is the covariance matrix, which is assumed to be diagonal. Then the log-likelihood of the sequence of purchases of all individuals $i \in \{1, \ldots, m\}$ for $t = \{1, \ldots, T_i\}$ is equal to $\sum_{i=1}^{m} \log \left( \prod_{t=1}^{T_i} \frac{\exp(I_{j_{it}} \beta_{j_{it}}^0 + \beta_{j_{it}})}{1 + \sum_{a_l \in S_{it}} \exp(I_{l_{it}} \beta_{l_{it}}^0 + \beta_{l_{it}})} \right) \phi(\beta) d\beta$ such that $I_{j_{it}} \beta_{j_{it}}^0 + \beta_{j_{it}}$ is a parameter value of product $j_{it} \in \{1, 2, \ldots, n\}$, where $I_{j_{it}} = 1$ if product $j_{it}$ is under promotion at time $t$ for individual $i$, and 0 otherwise. Model parameters are estimated through maximum simulated likelihood estimation (MSLE) where we use the simulated probabilities to approximate the following log-likelihood function:

$$
\max_{\mu, \Sigma} \sum_{i=1}^{m} \log \left( \prod_{t=1}^{T_i} \frac{\exp(I_{j_{it}} \beta_{j_{it}}^0 + \beta_{j_{it}})}{1 + \sum_{a_l \in S_{it}} \exp(I_{l_{it}} \beta_{l_{it}}^0 + \beta_{l_{it}})} \right) \phi(\beta) d\beta,
$$

where for any random draw $r = 1, 2, \ldots, R$ of a random vector $\xi^r$; that is sampled as $2n$-dimensional multivariate standard normal, we have that $\beta_{\ell}^r = \mu_{\ell} + \xi_{\ell}^r \sigma_{\ell}$, for any $\ell = 1, 2, \ldots, n$, and $\beta_{n+\ell}^r = \mu_{\ell} + \xi_{\ell}^r \sigma_{\ell}$, $\ell = n + 1, n + 2, \ldots, 2n$. The above optimization problem is nonconcave. To solve the problem we choose $R = 400$ and use a general nonlinear solver to converge to a stationary point (see Appendix A2.1.1 in Jagabathula and Vulcano 2017). Then we make predictions as follows:

$$
f(j_{it}, S_{it}) = \int \frac{\exp(I_{j_{it}} \beta_{j_{it}}^0 + \beta_{j_{it}})}{1 + \sum_{a_l \in S_{it}} \exp(I_{l_{it}} \beta_{l_{it}}^0 + \beta_{l_{it}})} \hat{\phi}(\beta|H_i; \mu, \Sigma) d\beta,$$

where $f(j_{it}, S_{it})$ is the probability to choose an item $j_{it}$ from the offer set $S_{it}$ for individual $i$, $\hat{\phi}(\beta|H_i; \mu, \Sigma)$ is the posterior distribution of parameter vector $\beta$ for customer $i$, conditioning on population prior and observed choices of customer $i$, i.e., $H_i = \{(a_{j_{it}}, S_{it}) : 1 \leq t \leq T_i\}$.

A3. Analyses of PO-Based Choice Models

In this section, we analyze the applicability of the heuristics to decycle the preference graph of individuals and illustrate the behavior and quality of the bounds we have developed in the present paper. To streamline our analysis and simplify the exposition, we ignore the information about product promotions by focusing on the most generic partial order based model which is PO-MNL Standard choice model proposed by Jagabathula and Vulcano (2017).

A3.1. Heuristics for Preference Graph Decycling

In the Section 2.3, we formulated a MIP (2) that decycles the preference graph of individuals with cycles. Since solving the MIP to optimality is challenging in general, we develop a tractable heuristics algorithm to decycle the preference graph. First, let $F(a_k, a_j, G)$ denote Dijkstra’s algorithm (polynomial in the number of nodes in the graph) to find the shortest path between nodes $a_k$ and
$a_j$ in the directed graph $G$, which returns the set of weighted edges comprising this path. To this end we propose an Algorithm 1 below (based on the Dijkstra’s algorithm $F(a_k,a_j,G)$) that provides a very good approximation of MIP (2) to decycle preference graphs with cycles.

**Algorithm 1** Preference graph $G$ decycling algorithm

1: procedure DECAYE($G$) \quad $\triangleright$ Where $G$ is a graph with the set of weighted edges $E_G$, and the set of nodes $V_G$

2: for $a_k$ in $V_G$ do

3: for $a_j$ in $V_G \setminus \{a_k\}$ do

4: while $F(a_k,a_j,G) \neq \emptyset$ & $F(a_j,a_k,G) \neq \emptyset$ do

5: Cycle ← $F_D(a_k,a_j,G) \cup F_D(a_j,a_k,G)$

6: $(a_x,a_y)$ ← the edge with the minimum weight in Cycle

7: $G'$ is obtained from $G$ such that $V_{G'} = V_G$, and $E_{G'} = E_G \setminus \{a_x,a_y\}$

8: $G \leftarrow G'$

9: return DAG $D = G$

Next we consider the empirical results of comparing MIP (2) with Algorithm 1 based on the actual sales dataset (see description of the sales data in the Section 4.1). We restrict the analysis only to individuals with cycles in the preference graph under the PO-MNL Standard choice model. Note that the DAG construction under the Standard model consists only of Phase 1 (i.e., based on the purchasing transactions, we build the preference graph with only candidate edges) and Phase 2 (i.e., we infer the DAG from the preference graph by deleting some of the candidate edges), since we ignore the information about promotions (see Section 2.3). First, when considering the top panels in the Figure A3 we observe that MIP (2) reduces the number of edges in the preference graph with cycles by 7.7% (deleting spurious edges) and reduces the weight of the preference graph by 5.02% on average across 27 product categories. In the bottom panels of the Figure A3 we represent the scatter plot over 27 product categories of the average number of edges in the DAG after applying MIP (2) versus after applying Algorithm 1, and another scatter plot over 27 product categories of the average weight of the DAG after applying MIP (2) versus after applying Algorithm 1. In particular, after applying approximation Algorithm 1 for the preference graph decycling we delete 0.16% more edges than after applying MIP (2), and the total weight of the DAG after applying Algorithm 1 is lower by 0.07% than after applying MIP (2), on average across 27 product categories.

**A3.2. Evaluation of Analytical Bounds**

The presence of $v$-nodes (i.e., nodes with more than one incoming edge) in DAGs of individuals complicates the maximum likelihood estimation of parameter values under PO-based choice models.
The left column of the panels in the Figure A4 illustrates that individuals without cycles in their preference graph have on average 0.04 v-nodes in their DAG, while individuals with cycles in their preference graph have on average 0.43 v-nodes in their DAG. Then the tractable approximation of the likelihood of a DAG $D$ is given by

$$\tilde{\lambda}(D) = \prod_{a_j \in \mathcal{N}} \sum_{a_k \in \Psi_D(a_j)} \frac{v_j}{v_k},$$

where $v_j = \exp(\beta_j), \forall a_j \in \mathcal{N}$ and $\Psi_D(a_j)$ denotes the reachability function such that $\Psi_D(a_j) = \{a_k : a_k \text{ is reachable from } a_j \text{ in } D\}$. Note that $\Psi_D(a_j)$ is always nonempty, since we assume that each node $a_j$ is reachable from itself. The approximation $\tilde{\lambda}(D)$ of the likelihood of DAG $D$ is exact when $D$ is a forest of directed trees, each with a unique root. We show in the Proposition 1 that $\tilde{\lambda}(D)$ is the lower bound of the DAG $D$’s likelihood.

Next in order to find the upper bound approximation of DAG $D$’s likelihood, let us denote $\bar{D}$ as the DAG obtained from $D$ where for every node with more than one incoming edge in DAG $D$ we delete all the incoming edges, except exactly the one. Instead of deleting an arbitrary set of edges, one can determine edges to delete to make the approximation as tight as possible. Finding the tightest upper bound is challenging in general. In order to ease the computational process, we develop a greedy-type heuristic $\Phi(D)$ (see Algorithm 2) to obtain the tightest upper bound of DAG $D$’s likelihood, i.e., $\bar{D} = \Phi(D)$ and $\lambda(D) \leq \lambda(\bar{D})$. Let $\log \mathcal{L}(X, \beta)$ denote the upper bound approximation of the panel data log-likelihood function under PO-based model s.t. $\log \mathcal{L}(X, \beta) =$

---

**Algorithm 2** DAG $D$ transformation to find its upper bound likelihood

```
1: procedure $\Phi(D)$  \(\triangleright\) Where $\Phi(D)$ is the DAG with each node having a unique parent s.t. $\lambda(\Phi(D)) \geq \lambda(D)$
2:     $A \leftarrow \mathcal{F}_D$                     \(\triangleright\) The set of nodes in $D$ with more than one incoming edge
3:     $B_i \leftarrow \{ (a_j, a_k) \in E_D : a_k = a_i \}$  \(\triangleright\) The set of incoming edges to node $a_i$
4: for $a_i \text{ in } \mathcal{F}_D$ do
5:     $D'$ is obtained from $D$: $V_{D'} = V_D$, and $E_{D'} = E_D \setminus B_i$
6:     $D \leftarrow D'$
7: while $A \neq \emptyset$ do
8:     $(a_x, a_y) = \arg\min_{(a_j, a_i) \in B_i} \lambda(D')$ s.t. $D': V_{D'} = V_D$, $E_{D'} = E_D \cup (a_j, a_i)$ and $a_i \in A$
9:     $D \leftarrow D'$
10:    $A \leftarrow A \setminus \{a_y\}$
11: return DAG $\bar{D} = \Phi(D)$  \(\triangleright\) Where $\bar{D}$ is the DAG with each node having a unique parent s.t. $\lambda(\bar{D})) \geq \lambda(D)$
```
\[ \sum_{i=1}^{m} \log(\lambda(D_i)), \text{ and let } \log \mathcal{L}(X, \beta) \text{ denote the lower bound approximation of the panel data log-likelihood function under PO-based model s.t. } \log \mathcal{L}(X, \beta) = \sum_{i=1}^{m} \log(\hat{\lambda}(D_i)). \] Then letting \( \beta \) be the solution to the maximization problem of the upper bound of the likelihood function, i.e.,

\[
\beta = \arg \max_{\beta^*} \log \mathcal{L}(X, \beta^*),
\]

we have that the maximum value of the exact log-likelihood function, i.e., \( \max_{\beta^*} \log \mathcal{L}(X, \beta^*) \), is upper bounded by \( \log \mathcal{L}(X, \beta^*) \), and lower bounded by \( \log \mathcal{L}(X, \beta) \):

\[
\log \mathcal{L}(X, \beta) \leq \log \mathcal{L}(X, \beta^*) \leq \max_{\beta^*} \log \mathcal{L}(X, \beta^*) = \log \mathcal{L}(X, \beta).
\]

Then the middle column of the panels in the Figure A4 illustrates that the upper bound of the log-likelihood function (i.e., \( \log \mathcal{L}(X, \beta) \)) is higher than the lower bound of the log-likelihood function (i.e., \( \log \mathcal{L}(X, \beta) \)) by 4.67\% for individuals without cycles in their preference graph, and by 13.39\% for individuals with cycles in their preference graph, on average across 27 product categories.

Next we illustrate the behavior and quality of the bounds we have developed for posterior probabilities of purchase when customers make choices consistently with their partial orders. In particular, we proposed the approximate probability of choosing product \( a_j \) from offer set \( S \) assuming that the sampled preference list is consistent with DAG \( D \):

\[
\hat{f}(a_j, S, D) = \begin{cases} 
\frac{\hat{\lambda}(D \cup C(a_j, S))}{\lambda(D)} & \text{if } a_j \in h_D(S), \\
0 & \text{otherwise}.
\end{cases}
\]

Letting \( \underline{f}(a_j, S, D) \) denote the lower bound of the purchase probability and \( \overline{f}(a_j, S, D) \) denote the upper bound of the purchase probability such that \( \underline{f}(a_j, S, D) = \frac{\hat{\lambda}(D \cup C(a_j, S))}{\lambda(D)} \) if \( a_j \in h_D(S) \) and 0, otherwise; \( \overline{f}(a_j, S, D) = \frac{\lambda(D \cup C(a_j, S))}{\lambda(D)} \) if \( a_j \in h_D(S) \), and 0, otherwise, we have the following inequalities:

\[
\underline{f}(a_j, S, D) \leq \hat{f}(a_j, S, D) \leq \overline{f}(a_j, S, D),
\]

\[
\underline{f}(a_j, S, D) \leq \underline{f}(a_j, S, D) \leq \overline{f}(a_j, S, D).
\]

The right column of Figure A4 illustrates that the percentage of transactions when prediction of the item to be chosen using the upper bound posterior probability of purchase, i.e., \( \overline{f}(a_j, S, D) \), is different from the prediction of the item to be chosen using the lower bound posterior probability of purchase, i.e., \( \underline{f}(a_j, S, D) \) is only 0.24\% for individuals without cycles in their preference graph and 0.37\% for individuals with cycles in their preference graph, where our prediction is the item with the highest probability of being purchased.

In the paper we use the following tractable formula to compute the posterior probabilities of purchase:

\[
\hat{f}(a_j, S, D) = \begin{cases} 
\frac{\Psi_D(a_j)}{\sum_{a_k \in h_D(S) \cup \Psi_D(a_k)}} & \text{if } a_j \in h_D(S), \\
0 & \text{otherwise,}
\end{cases}
\]
In the Figure A5 in the Appendix A4, we compare the choice prediction results made with $\tilde{f}(a_j, S, D)$ vs. the choice prediction results made with $\hat{f}(a_j, S, D)$ for individuals with and without cycles under “chi-square” score and miss rate (see description of the metrics in the Section 4.4). For all the panels in the Figure A5, the average MAE (Mean Absolute Error) is below 0.02% which indicates that posterior probability approximation $\tilde{f}(a_j, S, D)$ is very close to the posterior probability based on the lower bound of the DAG’s likelihood $\hat{f}(a_j, S, D)$ in terms of predictive performance.

A4. Figures

Figure A3  Top panels: The average percentage of the preference graph weight reduction and the average percentage of the preference graphs’ edges deleted after MIP (2) for every product category. Bottom panels: The scatter plot of the average weight of the DAG after MIP (2) vs. the average weight of the DAG after Algorithm 1, and the scatter plot of the average number of edges in the DAG after MIP (2) vs. the average number of edges in the DAG after Algorithm 1, over 27 product categories.
Figure A4  Top panels: Restrict the analysis to only individuals without cycles in the preference graph. Bottom panels: Restrict the analysis to only individuals with cycles in the preference graph. Left panels: The population average number of v-nodes for every product category. Middle panels: The population average lower and upper bounds of the negative DAG’s log-likelihood, i.e., $-\log \mathcal{L}(X, \beta)$ and $-\log \mathcal{L}(X, \beta)$ for every product category. Right panels: The percentage of transactions when prediction of the item to be chosen using the upper bound posterior probability of purchase, i.e., $\hat{f}(a_j, S, D)$, is different from the prediction of the item to be chosen using the lower bound posterior probability of purchase, i.e., $f(a_j, S, D)$, for every product category.

Figure A5  Left panels: Restrict the analysis to only individuals without cycles in the preference graph. Right panels: Restrict the analysis to only individuals with cycles in the preference graph. Top panels: The scatter plot over 27 product categories of the average $\chi^2$ score when prediction is based on $\hat{f}(a_j, S, D)$ vs. $\hat{f}(a_j, S, D)$. Bottom panels: The scatter plot over 27 product categories of the average miss rate when prediction is based on $\hat{f}(a_j, S, D)$ vs. $\hat{f}(a_j, S, D)$. 