A1. Preliminaries on DAGs

For completeness, we summarize the relevant notation from the main body and also introduce additional notation. Let \( \mathcal{N} = \{a_1, \ldots, a_n\} \) denote the universe of \( n \) products. For the purposes of this Appendix, we ignore the no purchase option and information about product promotions in order to simplify exposition. This assumption is without loss of generality since we can explicitly account for promotions by expanding the product universe and include the no purchase option as one more item. This is further developed in Section 5 in the main body of the paper.

A DAG \( D \) is a subset of pairwise preferences, \( \{(a_j, a_{j'}): 1 \leq j, j' \leq n\} \). We visualize a DAG \( D \) as a directed graph with nodes as products and a directed edge from \( a \) to \( b \) if the ordered pair \( (a, b) \in D \). 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 \( \mathcal{S}_n \) denote the collection of all possible \( n! \) rankings or permutations of the products in \( \mathcal{N} \). For any ranking \( \sigma \in \mathcal{S}_n \), we let \( \sigma(a) \) denote the preference ranking of product \( a \) under ranking \( \sigma \). We adopt the convention that lower ranked products are preferred over higher ranked ones, which means 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 \( S_D \) denote the subset of rankings that are consistent with \( D \); that is, \( S_D := \{\sigma \in \mathcal{S}_n : \sigma(a) < \sigma(b) \text{ whenever } (a, b) \in D\} \).

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 \( D \). Formally, \( \Psi_D(a) := \{b: \text{there is a directed path from } a \text{ to } b \text{ in } D\} \). We assume that \( a \) is reachable from itself, so
a ∈ Ψ(a). The set Θ_D(a) comprises the nodes from which a can be reached, i.e., Θ_D(a) := \{b: there is a directed path from b to a in D\}. To be consistent, we also include a in Θ_D(a). When the DAG D is clear from the context, we drop D from the notation and simply write Ψ(a) and Θ(a).

For any subset S ⊆ N, suppose π is a ranking of the products in S possibly including less than n products. Then, σ(π) denotes the set of all complete rankings of the products in N that are consistent with π, i.e., σ(π) = \{σ ∈ \mathcal{R}_n : σ(a) < σ(b) whenever π(a) < π(b)\}.

A2. Technical results

A2.1. Propositions and proofs in Section 2

A2.1.1. Preference graph decycling

Here we argue that the decycling procedure prioritizes retaining as many candidate edges as possible (measured by the sum of the weights involved). To that end, we show that the weight of the candidate edges in DAG D after preference graph G decycling is equal to the weight of the candidate edges in DAG D∗ such that D∗ ⊆ G has maximum weight of the candidate edges.

Let cw(D) denote the weight of the candidate edges in DAG D and let iw(D) denote the weight of the implicit candidate edges in DAG D. Let tw(D) denote the total weight of DAG D, i.e.,
tw(D) = cw(D) + iw(D).

Let us define D∗ as a DAG in G with the maximum weight of the candidate edges, i.e.,
D∗ = \arg\max\{cw(D) : D ⊆ G\}. Let D denote the DAG obtained from G after solving MIP (1). The next result follows.

**Proposition A1.** Candidate weight of DAG D, obtained after MIP (1) decycling applied over G is equal to the candidate weight of DAG D∗ such that D∗ = \max\{cw(D) : D ⊆ G\}, i.e.,
cw(D) = cw(D∗).

**Proof of Proposition A1:** Assume by contradiction that cw(D) < cw(D∗). Because cw(·) is always an integer, it follows that cw(D∗) − cw(D) ≥ 1. Further, note that D∗ is a feasible solution to the optimization problem MIP (1), which implies that tw(D) ≥ tw(D∗). It now follows from the definition of tw(·) that

\[ cw(D) + iw(D) ≥ cw(D∗) + iw(D∗) \implies iw(D) ≥ cw(D∗) − cw(D) + iw(D∗). \]

Because iw(D∗) ≥ 0 and cw(D∗) − cw(D) ≥ 1, we obtain that iw(D) ≥ 1. This, however, is not possible because of the scaling factor 1/(n^2T). Specifically, note that the maximum number of implicit candidate edges is n(n − 1) and the maximum possible weight of each implicit edge is \(T/(n^2T) = 1/n^2\). Therefore, the aggregate implicit weight is always bounded above by n(n − 1)/n^2, which is strictly less than 1. We have thus arrived at a contradiction. □
A2.1.2. Likelihood of the DAG-based choice model

**Proposition A2.** For a given set of parameters $\beta$ that characterize the distribution $\lambda$, the likelihood function of the DAG-based choice model is given by

$$\log L(\text{Panel Data}) = \sum_{i=1}^{m} \log \lambda(D_i) = \sum_{i=1}^{m} \log \left( \sum_{\sigma \in S_{D_i}} \lambda(\sigma) \right).$$

**Proof of Proposition A2:** For each individual $i$ and transaction $t$, let $C_{it} \subseteq S_{it}$ denote the consideration set of the individual. Because products outside the consideration set do not affect the individual’s choices, the data log-likelihood only depends on the choices and the consideration sets, rather than choices and offer sets. Letting $f(a_{jit}, C_{it}, D_i)$ denote the probability of purchasing product $a_{jit}$ from consideration set $C_{it}$ for individual $i$ with DAG $D_i$, we can write

$$\log L(\text{Panel Data}|\beta) = \sum_{i=1}^{m} \log \Pr \left( (a_{jit}, C_{it}) | t=1, D_i, \beta \right)$$

$$= \sum_{i=1}^{m} \log \left( \Pr(D_i|\beta) \cdot \Pr \left( (a_{jit}, C_{it}) | t=1, D_i, \beta \right) \right)$$

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

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

(A1)

The second equality follows from a straightforward application of the conditional probability formula. The third equality follows because conditioning on the DAG $D_i$, individual $i$’s purchase probabilities can be computed independently.

We now focus on the term $f(a_{jit}, C_{it}, D_i)$. Note that we only observe the offer sets $S_{it}$. The consideration sets $C_{it}$ are latent. Nevertheless, given $D_i$, we can constrain $C_{it}$ sufficiently to allow for the computation of $f(a_{jit}, C_{it}, D_i)$. There are two cases. First, we consider the case when none of the edges in the set $\{(a_{jit}, a_k) : a_k \in S_{it} \setminus \{a_{jit}\}\}$ was deleted in the decycling step. In this case, the customer always prefers product $a_{jit}$ over all the other products in the offer set $S_{it}$, and therefore, chooses product $a_{jit}$ irrespective of the consideration set. This implies that $f(a_{jit}, C_{it}, D_i) = 1$ for all $C_{it} \subseteq S_{it}$ such that $a_{jit} \in C_{it}$.

Next, we consider the case when some of the edges in the set $\{(a_{jit}, a_k) : a_k \in S_{it} \setminus \{a_{jit}\}\}$ are deleted in the decycling step. Because the decycling procedure deletes the smallest possible weight of edges, the edge $(a_{jit}, a_k)$ for some $a_k \in S_{it} \setminus \{a_{jit}\}$ is deleted only if there is a directed path from $a_k$ to $a_{jit}$ in the final DAG $D_i$. Now, because $a_k$ is preferred over $a_{jit}$, the only way the customer would choose $a_{jit}$ when $a_k$ was also on offer is if $a_k$ was not considered. We can thus conclude that
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; see Figure A1. 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 $v_{\Psi_D(a_y)} = \sum_{a_t \in \Psi_D(a_y)} v_t$. 

**Lemma A1.** Consider two subsets $S_1, S_2 \subset \mathcal{N}$ with $S_1 \cap S_2 = \emptyset$. Let $\pi_1$ be a ranking over $S_1$, and $\pi_2$ be a ranking over $S_2$. Assume w.l.o.g. that $\pi_1 = (a_{1,1}, a_{1,2}, \ldots, a_{1,k_1})$ and $\pi_2 = (a_{2,1}, a_{2,2}, \ldots, a_{2,k_2})$. For a fixed $i$ with $0 \leq i \leq k_1$, let $S_i(\pi_1, \pi_2)$ be the set of rankings in $\mathcal{S}_n$ 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.,

$$S_i(\pi_1, \pi_2) = \{\sigma \in \mathcal{S}_n : \sigma \in \sigma(\pi_1) \cap \sigma(\pi_2), \text{ with } \sigma(2,1) \geq i + 1\}.$$

Then,

$$\lambda(S_i(\pi_1, \pi_2)) = \prod_{j=1}^{i} \frac{v_{1,j}}{\sum_{j'=j}^{k_1} v_{1,j'}} \prod_{j=i+1}^{k_1} \frac{v_{1,j}}{\sum_{j'=j}^{k_1} v_{1,j'}} \lambda(\pi_2).$$

The first new result refers to an expression for the probability of DAG $D$ that is split in two independent terms by breaking the bottom part of $D$ in two pieces.
Figure A1 Illustration for the proof in Lemma A1. The bottom part of a DAG is split in two independent terms. Note that DAG \( D[a_y] \) is defined by nodes in \( \Psi_D(a_y) \).

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 \( N \setminus \Psi_D(a_y) \) and \( S_2 \) denote \( \Psi_D(a_y) \). It is clear that \( S_1 \cap S_2 = \emptyset \) and \( S_1 \cup S_2 = N \). 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 establish the following result.

Claim: The set \( E_D \) of edges in the transitive reduction of \( D \) can be partitioned as

\[
E_D = E_1 \cup E_2 \cup E_3, \quad \text{where} \quad E_3 = \{(a, a_y) : (a, a_y) \in E_D \} \quad \text{and} \quad E_i \cap E_j = \emptyset \quad \forall \ 1 \leq i \neq j \leq 3 \quad (A2)
\]

Proof. 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. Consider the following definitions:

- Let $\pi_1$ (of length $k_1$) and $\pi_2$ (of length $k_2$) be rankings of products in the sets $S_1$ and $S_2$, respectively. Note that $k_1 + k_2 = n$.
- 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 $\mathcal{S}_n$ 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., $S_i(\pi_1, \pi_2) = \{\sigma \in \mathcal{S}_n : \sigma \cap \sigma(\pi_2) = \sigma(\pi_2^{-1}(1)) \geq \sigma(\pi_1^{-1}(i)) + 1\}$.
- 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., $i(\pi_1) = \max \{\pi_1(a) : a \in \Theta_D(a_y)\}$.

**Claim:** 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

$$S_D = \bigcup_{(\pi_1, \pi_2) \in X} S_i(\pi_1, \pi_2), \text{ where } S_i(\pi_1, \pi_2) \cap S_i'(\pi_1', \pi_2') = \emptyset \text{ for } (\pi_1, \pi_2) \neq (\pi_1', \pi_2') \quad (A3)$$

**Proof.** We first show that $S_i(\pi_1, \pi_2) \subseteq S_D$ for all $(\pi_1, \pi_2) \in X$. For that, consider $\sigma \in S_i(\pi_1, \pi_2)$ and consider an edge $(a, b) \in E_D$. It is sufficient to show that $\sigma(a) < \sigma(b)$. It follows from $(A2)$ 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 $\sigma(a) = \pi_1(a) \leq \pi_1(a^*) = \sigma(a^*)$, with equality when $a = a^*$. Note that both rankings $\sigma$ and $\pi_1$ must coincide until position $i(\pi_1)$. 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 $\sigma(a) \leq \sigma(a^*) < \sigma(a_y) = \sigma(b)$, 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 $\sigma \in S_D$ and let $\pi_1$ and $\pi_2$ denote the rankings $\sigma$ induced 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 $\sigma \in S(\pi_1) \cap S(\pi_2)$. 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 $\sigma(a) < \sigma(a_y)$ 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 \( S_D = \bigcup_{(\pi_1, \pi_2) \in X} S_i(\pi_1, \pi_2) \). The disjointness of \( S_i(\pi_1, \pi_2) \) and \( S_i(\pi'_1, \pi'_2) \) 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 (A3).

We can then write from (A3) that

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

Now consider \( (\pi_1, \pi_2) \in X \). Without loss of generality, suppose that \( \pi_1 = (a_1, \ldots, a_{k_1}) \) and \( \pi_2 = (a_2, a_3, \ldots, a_{k_2}) \). Then, invoking (Jagabathula and Vulcano 2018, Lemma A1), we can write

\[
\lambda(S_i(\pi_1, \pi_2)) = \left[ \prod_{j=1}^{i(\pi_1)} \frac{v_{1,j}}{\sum_{j'=j}^{k_1} v_{1,j'} + \sum_{j'=1}^{k_2} v_{2,j'}} \right] \cdot \lambda(\pi_2)
\]

where the second equality follows from the fact that \( S_2 = \Psi_D(a_y) \), and where we define

\[
g(\pi_1) = \prod_{j=1}^{i(\pi_1)} \frac{v_{1,j}}{\sum_{j'=j}^{k_1} v_{1,j'} + \sum_{j'=1}^{k_2} v_{\Psi_D(a_y)}}
\]

We now have

\[
\lambda(D) = \sum_{\pi_1: \sigma(\pi_1) \subseteq S_D} \sum_{\pi_2: \sigma(\pi_2) \subseteq S_D} \lambda(S_i(\pi_1, \pi_2)) = \left[ \sum_{\pi_1: \sigma(\pi_1) \subseteq S_D} g(\pi_1) \right] \cdot \left[ \sum_{\pi_2: \sigma(\pi_2) \subseteq S_D} \lambda(\pi_2) \right].
\]

Noting that

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

we have shown that

\[
\lambda(D) = \lambda(D[a_y]) \cdot \left[ \sum_{\pi_1: \sigma(\pi_1) \subseteq S_D} g(\pi_1) \right]. \tag{A4}
\]

It now suffices to show that

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

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 \( D[a_y] \) with the nodes of the DAG decomposed into sets \( S_1 \) 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 (A4), we can write

$$\lambda_y(\bar{D}[a_y]) = \lambda_y(D_3) \cdot \sum_{\pi_1: \sigma(\pi_1) \subseteq S_{D_1}} g(\pi_1),$$

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

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

The result of the lemma now follows. $\square$

To establish the next result, we first quote the so-called internal consistency property referred to in Jagabathula and Vulcano (2018), i.e., the probability of a ranking in a subset of products only depends on the relative order of the elements in the subset.

**Proposition A1.2 in Jagabathula and Vulcano (2018).** For any subset $S$ of $k$ products, let $\pi$ be a ranking over the elements of $S$. Let $\sigma(\pi)$ denote the set of full rankings over $N$ defined as

$$\sigma(\pi) = \{\sigma: \sigma(a_i) < \sigma(a_j) \text{ whenever } \pi(a_i) < \pi(a_j) \text{ for all } a_i, a_j \in S\}.$$

Then, it must hold that

$$\lambda(\pi) = \sum_{\sigma \in \sigma(\pi)} \lambda(\sigma) = \prod_{r=1}^{k} \frac{v_{\pi_r}}{\sum_{j=r}^{k} v_{\pi_j}}.$$

We also need the following notation. Given a DAG $D$ and product $a_y$, let $\lambda_y^{\text{aug}}$ 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 $a'_y$ also 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 $\sum_{a_j \text{ is a } v\text{-node}} (d_{j}^n - 1)$, where $d_{j}^n$ 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

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

The inequality is strict when all the parameters under the MNL model are positive.

Furthermore the approximate likelihoods of the DAGs $D$ and $D^{\text{split}}$ are equal, i.e., $\tilde{\lambda}_y^{\text{aug}}(D^{\text{split}}) = \tilde{\lambda}(D) = \prod_{a \in N} \frac{v_a}{\sum_{a' \in \Psi_D(a)} v_{a'}}.$
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 $D_{split}$ 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, $E_{D_{split}} = E_D \setminus \{(a_1,a_y)\} \cup \{(a_2,a_y')\}$. Figure A2 illustrates these DAGs. Note that by construction, the $v$-degree of $D_{split}$ is one less than that of $D$ because the in-degree of node $a_y$ has been reduced by 1.

We need the following intermediate result.

Claim: $\lambda(D) = \lambda_{aug}(D_1)$.

Proof. Note that since $a_y$ is a leaf node in $D$, we have that $v_{\Psi_D}(a_y) = v_y$. Therefore, the distribution $\lambda_{aug}$ is defined on the expanded universe $N \cup \{a_y'\}$ with the weights of the products in $N$ remaining the same as in $\lambda$ and the weight $v_y$ assigned to product $a_y'$.

Now, for any ranking $\pi \in S_D$ (including only products in set $N$), let $\sigma(\pi) \subset S_{D_1}$ denote the set of rankings of the products in the set $N \cup \{a_y'\}$ that are consistent with $\pi$. By invoking (Jagabathula and Vulcano 2018, Proposition A1.2), it holds that $\lambda(\pi) = \sum_{\sigma \in \sigma(\pi)} \lambda_{aug}(\sigma)$. We can now write

$$\lambda_{aug}(D_1) = \sum_{\sigma \in S_{D_1}} \lambda_{aug}(\sigma) = \sum_{\pi \in S_D} \sum_{\sigma \in \sigma(\pi)} \lambda_{aug}(\sigma) = \sum_{\pi \in S_D} \lambda(\pi) = \lambda(D).$$

Therefore, in order to establish that $\lambda_{aug}(D_{split}) \leq \lambda(D)$, it is sufficient to show that $\lambda_{aug}(D_{split}) \leq \lambda_{aug}(D_1)$. For that, consider the three DAGs $I_1$, $I_2$, and $I_3$ (see Figure A3), defined over the set of products in $N \cup \{a_y'\}$ such that

$E_{I_1} = E_{D_1} \cup \{(a_2,a_y')\}$, $E_{I_2} = E_{D_1} \cup \{(a_y',a_2)\}$, 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

$$\lambda_{aug}(D_1) = \lambda_{aug}(I_1) + \lambda_{aug}(I_2)$$

$$\lambda_{aug}(D_{split}) = \lambda_{aug}(I_1) + \lambda_{aug}(I_3).$$
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 \( h : I_3 \to I_2 \) 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 \( \sigma' = (\ldots, a_1, \ldots, a_y', \ldots, a_2, \ldots, a_y') \) 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., \( v_{a_y} = v_{a_y'} \), 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

\[
\lambda_y^{\text{aug}}(I_3) = \sum_{\sigma \in I_3} \lambda_y^{\text{aug}}(\sigma) = \sum_{\sigma \in I_3} \lambda_y^{\text{aug}}(h(\sigma)) \leq \sum_{\sigma' \in I_2} \lambda_y^{\text{aug}}(\sigma') = \lambda_y^{\text{aug}}(I_2),
\]

(A5)

where the inequality holds because there could be additional rankings \( \sigma' \in I_2 \) that do not have a pre-image in \( I_3 \) through \( h(\cdot) \). We have thus shown that \( \lambda_y^{\text{aug}}(D^{\text{split}}) \leq \lambda_y^{\text{aug}}(D_1) \), which implies that \( \lambda_y^{\text{aug}}(D^{\text{split}}) \leq \lambda(D) \).

The inequality in \( \lambda_y^{\text{aug}}(D^{\text{split}}) \leq \lambda(D) \) is strict when all the parameters under the MNL model are positive. We establish this result by showing that there exists \( \sigma' \in I_2 \) such that \( h(\sigma) \neq \sigma' \) for all \( \sigma \in I_3 \). It then follows that the inequality in (A5) is strict, implying that \( \lambda_y^{\text{aug}}(D^{\text{split}}) < \lambda(D) \).

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 \( \sigma = (\ldots, a_1, \ldots, a_y, \ldots, a_2, \ldots, a_y', \ldots) \), 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 \( \tilde{\lambda}_y^{\text{aug}}(D^{\text{split}}) = \hat{\lambda}(D) \). 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 \( \lambda_y^{\text{aug}} \), are equal by definition, and the approximations \( \tilde{\lambda} \) and \( \tilde{\lambda}_y^{\text{aug}} \) only depend on the reachability weights, then the equality \( \tilde{\lambda}_y^{\text{aug}}(D^{\text{split}}) = \hat{\lambda}(D) \) immediately follows.

\( \square \)

We can now proceed to prove the results in Section 3.
Proof of Proposition 1

We show the result, $\lambda(D) \leq \lambda(D)$, 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. Then, $D$ is a forest of directed trees each with a unique root. It follows from (Jagabathula and Vulcano 2018, Proposition 3.2) that $\tilde{\lambda}(D) = \lambda(D) = \prod_{a \in \mathcal{N}} \frac{v_a}{\sum_{a' \in \Psi_D(a)} v_a}$, 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 with unique root. 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

$$
\tilde{\lambda}(D) = \prod_{j \in \mathcal{N}} \frac{v_j}{\sum_{j' \in \Psi_D(a_j)} v_{j'}}
$$

$$
= \left( \prod_{j \in \Psi_D(a_y)} \frac{v_j}{\sum_{j' \in \Psi_D(a_j)} v_{j'}} \right) \cdot \left( \prod_{j \in \mathcal{N} \setminus \Psi_D(a_y)} \frac{v_j}{\sum_{j' \in \Psi_D(a_j)} v_{j'}} \right)
$$

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

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

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 2018, Proposition 3.2), and the fact that $\Psi_D(a_j) = \Psi_{\bar{D}[a_y]}(a_j)$ for all $j \in \mathcal{N}$ such that $a_y \notin \Psi_D(a_j)$. We now have

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

$$
= \lambda(D[a_y]) \cdot \tilde{\lambda}_y(\bar{D}[a_y]) \cdot \lambda_y^{\text{aug}}(D_y^{\text{split}})
$$

$$
\leq \lambda(D[a_y]) \cdot \lambda_y^{\text{aug}}(D_y^{\text{split}}) \quad \text{[by the induction hypothesis]}
$$

$$
\leq \lambda(D[a_y]) \cdot \lambda_y(\bar{D}[a_y]), \text{ with strict inequality if } v_j > 0 \forall a_j \in \mathcal{N} \quad \text{[by Lemma A2]}
$$

$$
= \lambda(D) \quad \text{[by Lemma A1]},
$$
where the second equality holds by Lemma A2 taking $\bar{D}[a_y]$ here as $D$ there, and $D_y^{split}$ here as $D^{split}$ there; and the first inequality follows from induction hypothesis with distribution $\mu = \lambda^{aug}$ since the $v$-degree of $D_y^{split}$ is equal to $p$. The result of the proposition now follows. \hfill \Box

Proof of Proposition 2
We must have that $S_D \subset S_{\bar{D}}$ since if $\sigma$ is consistent with $D$, i.e., $\sigma \in S_D$, then it must also be consistent with $\bar{D}$, i.e., $\sigma \in S_{\bar{D}}$. It now follows that

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

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

$$\lambda(\bar{D}) - \lambda(D) \geq \lambda(\sigma) > 0,$$

which holds because all the parameters of $\lambda$ are positive. \hfill \Box

For the proof of Proposition 3 we recall here some notation. Let $R(D, \bar{D})$ denote the ratio of the upper bound to the lower bound $\lambda(\bar{D})/\tilde{\lambda}(D)$, for any DAG $\bar{D} \subset D$. Let $\ell$ denote the size of the largest reachability set in DAG $D$, i.e., $\ell = \max_{a \in N} |\Psi_D(a)|$, and let $p$ denote the number of nodes with $v$-nodes in their reachability sets, i.e., $p = \{|a \in N: \exists v\text{-node } b \in \Psi_D(a)\}$. Further, let $\Delta := \max_{a} \max_{b \in \Psi_D(a) \setminus \{a\}} \frac{v_b}{v_a}$ be the maximum ratio between the weights of nodes within the same directed path in the DAG.

Proof of Proposition 3
Define $\Phi(D) \subset D$ as a DAG with each node having a unique parent, such that for any distribution $\lambda$, $\lambda(\Phi(D)) \geq \lambda(D)$. Recall that $\bar{D}$ is a forest of directed trees obtained by deleting arcs from $D$ in order to break the $v$-nodes. Thus, set $\bar{D} = \Phi(D)$ so that each node has a unique parent, verifying

$$\lambda(\bar{D}) = \prod_{a \in N} \frac{v_a}{\sum_{a_j \in \Psi_{\Phi(D)}(a)} v_j}.$$ 

In turn, $\tilde{\lambda}(D)$ is the lower bound obtained by treating $D$ as a forest of directed trees with unique root. That is,

$$\tilde{\lambda}(D) = \prod_{a \in N} \frac{v_a}{\sum_{a_j \in \Psi_D(a)} v_j}.$$
We have from Propositions 1 and 2:

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

(A6)

Then,

\[
\log R(D, \tilde{D}) = \log \left( \frac{\lambda(\tilde{D})}{\lambda(D)} \right) = \log \left( \prod_{a \in N} \frac{\sum_{j \in \Psi_D(a)} v_j}{\sum_{j \in \Phi(D)(a)} v_j} \right)
\]

\[
= \sum_{a \in N} \log \left( 1 + \frac{\sum_{j \in \Psi_D(a) \setminus \Phi(D)(a)} v_j}{\sum_{j \in \Phi(D)(a)} v_j} \right)
\]

\[
= \sum_{a \in N} \|F_D \cap \Psi_D(a) \neq \emptyset \| \cdot \log \left( 1 + \frac{\sum_{j \in \Psi_D(a) \setminus \Phi(D)(a)} v_j}{\sum_{j \in \Phi(D)(a)} v_j} \right)
\]

\[
\leq \sum_{a \in N} \|F_D \cap \Psi_D(a) \neq \emptyset \| \cdot \log \left( 1 + \sum_{j \in \Psi_D(a) \setminus \Phi(D)(a)} \frac{v_j}{v_a} \right)
\]

\[
\leq \sum_{a \in N} \|F_D \cap \Psi_D(a) \neq \emptyset \| \cdot \log \left( 1 + \sum_{j \in \Psi_D(a) \setminus \Phi(D)(a)} \Delta \right)
\]

\[
\leq \sum_{a \in N} \|F_D \cap \Psi_D(a) \neq \emptyset \| \cdot \log(1 + \ell \cdot \Delta) \leq p \cdot \log(1 + \ell \cdot \Delta),
\]

where the fourth equality follows since the surviving terms are the ones where \(\Psi_D(a) \setminus \Phi(D)(a) \neq \emptyset\), i.e., there are nodes in \(\Psi_D(a)\) with more than one incoming edge; and the first inequality holds because \(a \in \Psi(a)\), and we add terms in the numerator and take out terms from the denominator.

The last three inequalities follow from the definitions of \(\Delta\), \(\ell\), and \(p\), respectively.

From (A6), we have that

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

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

since as \(n \to \infty\), it can be shown that \(n \Delta \in o(n^{-1})\) and \(n^2 \Delta \in o(1)\).
Proof of Proposition 4

Recall that the merged DAG $D \cup C(a_j, S)$ is obtained by taking the union of the graphs $D$ and $C(a_j, S)$. From the definitions of the bounds in (4), it must hold that

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

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

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

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 $\Delta n^2 \in o(1)$, as $n \to \infty$. \qed

A3. Heuristic for preference graph decycling

In Section 2.4 we formulated MILP (1) for preference graph decycling (Phase 3) in the DAG construction process. Since solving the MILP to optimality could be challenging (e.g., if we have thousands of products or brands), we propose a tractable, greedy heuristic to decycle the preference graph.

Let $\text{FindPath}(a_k, a_j, G)$ denote the output of Dijkstra’s algorithm on a directed graph $G$, which finds the shortest path between nodes $a_k$ and $a_j$ and returns the set of weighted edges comprising this path (potentially, the empty set). The Dijkstra’s algorithm runs in $O(|V_G|^2)$ time, where $V_G$ is the set of nodes in $G$.

Taking advantage of the polynomial running time of Dijkstra’s, our heuristic proceeds as follows: For a directed graph $G$, we run Dijkstra’s between all pair of nodes $a_k$ and $a_j$, in both directions. In case both paths exist, then there is a cycle containing $a_k$ and $a_j$, and the edge with minimum weight is removed. As a result, since there are $O(|V_G|^2)$ pairs of nodes, the preference graph decycling can be implemented with $O(|V_G|^4)$ computational complexity. The steps are described in Algorithm 1 below.

In order to validate the effectiveness of Algorithm 1, next we compare empirical prediction results on the actual sales dataset obtained from DAGs decyled via MILP (1) and the analogous results from DAGs decyled via Algorithm 1. A description of the sales data is provided in Section 4.1 in the main body of the paper.

In Figure A4 (left panels), we observe that using MILP (1) we delete 6.27% fewer edges than when using Algorithm 1 and obtain 0.4% denser DAGs on average across 27 product categories. In the middle and right panels we represent the scatter plot over 27 product categories of the average miss rate and $\chi^2$ scores, respectively. It follows that by using MILP (1) we obtain 1.17% lower miss
Algorithm 1 Preference graph $G$ decycling

1: \textbf{procedure} \textsc{Decycle}(G) \Comment{Where $G$ is a graph with set of nodes $V_G$ and set of weighted edges $E_G$.}
2: \textbf{for} $a_k$ in $V_G$ \textbf{do}
3: \hspace{1em} \textbf{for} $a_j$ in $V_G \setminus \{a_k\}$ \textbf{do}
4: \hspace{2em} \textbf{while} $\text{FindPath}(a_k, a_j, G) \neq \emptyset \& \text{FindPath}(a_j, a_k, G) \neq \emptyset$ \textbf{do}
5: \hspace{3em} $\text{Cycle} \leftarrow \text{FindPath}(a_k, a_j, G) \cup \text{FindPath}(a_j, a_k, G)$
6: \hspace{2em} Remove the edge $(a_x, a_y)$ with minimum weight in $\text{Cycle}$ from the set $E_G$
7: \textbf{return} DAG $D = G$

rate and 2.75% lower $\chi^2$ score than by using Algorithm 1. These results provide good support for the use of the greedy heuristic as an alternative to the exact solution of MIP (1) in cases where the number of products is large. We highlight here though that in all our experiments reported in the main body of the paper we used MIP (1) limited to a max time of 30 seconds, and retaining the best feasible solution when optimality was not reached. According to Table 1, the largest category contains 95 products in our case.

Figure A4 Comparison of the performance of heuristic Algorithm 1 vs. MIP (1) for 27 product categories. Left panel: scatter plot of the average percentage of edges deleted. Middle panel: scatter plot of the average miss rate. Right panel: scatter plot of the average $\chi^2$ score.

A4. Benchmark models

A4.1. 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 $h \in \{1, \ldots, K\}$ make choices according to the single class MNL model with a parameter value $\beta_{0j}^h + I_{jit} \beta_{jit}$ of product $j_{it} \in \{1, 2, \ldots, n\}$, where $I_{jit} = 1$ if product $j_{it}$ 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:

$$\max_{\beta, \gamma: \beta_h^{0} = 0, h \in [K]} \sum_{i=1}^{m} \log \left( \sum_{h=1}^{K} \gamma_h \prod_{t=1}^{T_i} \frac{\exp(\beta_{ht}^0 + I_{jit} \beta_{ht})}{\sum_{a \in S_{it}} \exp(\beta_{a}^0 + I_{ait} \beta_a)} \right) - \alpha \sum_{h=1}^{K} (\| \beta_h^0 \|_1 + \| \beta_h \|_1)$$

When the value of $\alpha$ is fixed and $K = 1$, it can be shown that this optimization problem is globally concave and therefore can be solved efficiently (Train 2009). Note that we tuned 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 2018). Specifically, we initialize the EM with a random allocation of customers to one of the $K$ classes, resulting in an initial allocation $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$, which form a partition of the collection of all the customers. Then we set $\gamma_h^{(0)} = |\mathcal{D}_h| / \left( \sum_{d=1}^{K} |\mathcal{D}_d| \right)$. In order to get a parameter vector $(\beta_0^{(0)}, \beta_h)$, we fit a single class MNL model to each subset of customers. Based on each customer $i'$s purchase history $(a_{jit}, S_{it})$ for $1 \leq t \leq T_i$, we can estimate their posterior membership probabilities $\forall h \in \{1, \ldots, K\}$:

$$\hat{\gamma}_{ih} = \frac{\gamma_h \prod_{t \in T_i} \left[ \exp(\beta_{hjit}^0 + I_{jit} \beta_{hjt}) / \left( \sum_{a \in S_{it}} \exp(\beta_{hat}^0 + I_{ait} \beta_{ht}) \right) \right]}{\sum_{d=1}^{K} \gamma_d \prod_{t \in T_i} \left[ \exp(\beta_{djit}^0 + I_{jit} \beta_{djt}) / \left( \sum_{a \in S_{it}} \exp(\beta_{hat}^0 + I_{ait} \beta_{ht}) \right) \right]},$$

and the prediction can be made as follows:

$$f(j_{it}, S_{it}) = \sum_{h=1}^{K} \frac{\hat{\gamma}_{ih} \exp(\beta_{hjit}^0 + I_{jit} \beta_{hjt})}{\sum_{a \in S_{it}} \exp(\beta_{hat}^0 + I_{ait} \beta_{ht})},$$

where $f(j_{it}, S_{it})$ is a probability to choose an item $j_{it}$ from the offer set $S_{it}$.

A4.2. 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 $\sum_{i=1}^{m} \log \left( \prod_{t=1}^{T_i} \frac{\exp(\beta_{jit}^0 + I_{jit} \beta_{jit})}{\sum_{a \in S_{it}} \exp(\beta_{a}^0 + I_{ait} \beta_a)} \right)$ such that $\beta_{jit}^0 + I_{jit} \beta_{jit}$ is a parameter value of product $j_{it} \in \{1, 2, \ldots, n\}$, where $I_{jit} = 1$ 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:

$$\max_{\mu, \Sigma: \mu_1 = 0} \sum_{i=1}^{m} \log \left( \prod_{t=1}^{T_i} \frac{\exp(\beta_{jir}^0 + I_{jirt} \beta_{jir})}{\sum_{a \in S_{it}} \exp(\beta_{a}^0 + I_{ait} \beta_a)} \right).$$
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 \( \beta^r_\ell = \mu_\ell + \xi^r_\ell \sigma_\ell \), for any \( \ell = 1, 2, \ldots, n \), and \( \beta^0_{r_{n+\ell}} = \mu_{n+\ell} + \xi^r_{n+\ell} \sigma_{n+\ell} \), \( \ell = n+1, n+2, \ldots, 2n \). 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 2018). Then we make predictions as follows:

\[
f(\mathbf{j}_{it}, S_{it}) = \int \frac{\exp(\beta^0_{j_{it}} + I_{j_{it}} \beta_{j_{it}})}{\sum_{a_j \in S_{it}} \exp(\beta^0_{a_j} + I_{a_j} \beta_{a_j})} \hat{\phi}(\beta|H_i; \mu, \Sigma) d\beta,
\]

where \( f(\mathbf{j}_{it}, S_{it}) \) 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., \( H_i = \{(a_{j_{it}}, S_{it}) : 1 \leq t \leq T_i\} \).

### A5. Evaluation of analytical bounds

In this section, we focus on the PO-MNL Promotion model and illustrate the behavior and quality of the bounds proposed in Section 3.

#### A5.1. Bounds on the probability of a DAG

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 panels in Figure A5 illustrate that individuals without cycles in their preference graph have on average 17.09 \( v \)-nodes in their DAG whereas individuals with cycles in their preference graph have on average 19.93 \( v \)-nodes in their DAG.

A tractable approximation of the likelihood of a DAG \( D \) is given by

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

where \( v_j = \exp(\beta_j), \forall a_j \in \mathcal{N} \) 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 \( \hat{\lambda}(D) \) of the likelihood of DAG \( D \) is exact when \( D \) is a forest of directed trees, each with a unique root. We show in Proposition 1 that \( \hat{\lambda}(D) \) is a lower bound for the likelihood of DAG \( D \).

Next, in order to find the upper bound approximation of DAG \( D \) likelihood, let us denote \( \bar{D} \) the DAG obtained from \( D \) where for every node with more than one incoming edge we delete all the incoming edges but one. Instead of deleting an arbitrary set of edges, we 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
Figure A5  Analysis of bounds for the probability of a DAG. Top panels: Analysis restricted to individuals without cycles in their preference graph. Bottom panels: Analysis restricted to individuals with cycles in their preference graph. Left panels: Population average number of v-nodes for every product category. Middle panels: Population average lower and upper bounds of the negative of the DAG log-likelihood, i.e., $-\log L(X, \beta)$ and $-\log \bar{L}(X, \beta)$, for every product category. Right panels: Percentage of transactions when prediction of the item to be chosen using the upper bound posterior probability of purchase, i.e., $f(a_j, S, D)$, is different from the prediction of the item to be chosen using the lower bound posterior probability of purchase, i.e., $\bar{f}(a_j, S, D)$, for every product category.

heuristic $\Phi(D)$ (see Algorithm 2) to obtain a tight upper bound of DAG $D$ likelihood, i.e., $\bar{D} = \Phi(D)$ and $\lambda(D) \leq \lambda(\bar{D})$. 
Algorithm 2: DAG $D$ transformation to find its upper bound likelihood

1: procedure $\Phi(D)$, \hspace{1em} $\triangleright$ where $\Phi(D)$ is the DAG with each node having a unique parent s.t. $\lambda(\Phi(D)) \geq \lambda(D)$
2: \hspace{1em} $A \leftarrow F_D$ \hspace{1em} $\triangleright$ $F_D$ is the set of nodes in $D$ with more than one incoming edge
3: \hspace{1em} for $a_i$ in $F_D$ do
4: \hspace{2em} $D'$ is obtained from $D$: $V_{D'} = V_D$, and $E_{D'} = E_D \setminus B_i$, \hspace{1em} $\triangleright$ where $B_i$ is the set of incoming edges into node $a_i$
5: \hspace{1em} $D \leftarrow D'$
6: \hspace{1em} while $A \neq \emptyset$ do
7: \hspace{2em} $(a_x, a_y) = \arg\min_{(a_i, a_j) \in B_i} \lambda(D')$ s.t. $D': V_{D'} = V_D$, $E_{D'} = E_D \cup (a_i, a_j)$ and $a_j \in A$
8: \hspace{2em} $D \leftarrow D'$
9: \hspace{2em} $A \leftarrow A \setminus \{a_y\}$
10: return DAG $\bar{D} = \Phi(D)$

For a collection of panel data represented by $X$ and a given set of parameters $\beta$, let $\log L(X, \beta)$ denote the upper bound approximation of the log-likelihood function under PO-MNL Promotion model defined as $\log L(X, \beta) = \sum_{i=1}^m \log \lambda(\bar{D}_i)$. Then, letting $\bar{\beta}^*$ be the solution to the maximization problem of the upper bound of the likelihood function, i.e., $\bar{\beta}^* = \arg\max_{\beta} \log L(X, \beta)$, we have that the maximum value of the exact log-likelihood function $\log L(X, \beta^*)$ satisfies:

$$\log L(X, \beta^*) \leq \log L(X, \bar{\beta}^*) \leq \log L(X, \beta^*) .$$

Similarly, let $\log \underline{L}(X, \beta) = \sum_{i=1}^m \log \tilde{\lambda}(D_i)$ denote the lower bound approximation of the log-likelihood function under PO-MNL Promotion model, with optimal values $\underline{\beta}^*$. Then,

$$\log \underline{L}(X, \underline{\beta}^*) \leq \log \underline{L}(X, \beta^*) \leq \log L(X, \beta^*) .$$

A natural question that arises is about the size of the gap between both easy-to-compute bounds (lower and upper). The middle column of the panels in Figure A5 illustrates that the upper bound of the log-likelihood function (i.e., $\log L(X, \beta)$) is higher than the lower bound of the log-likelihood function (i.e., $\log \underline{L}(X, \beta)$) by 4.72% for individuals without cycles in their preference graph, and by 6.79% for individuals with cycles in their preference graph, on average across 27 product categories. This observation provides good support to use any of the bounds as an approximation for the estimation problem under the exact likelihood of the DAGs. In particular, we used the lower bound $\log \underline{L}(X, \underline{\beta}^*)$.

A5.2. Bounds on the probability of purchase

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 propose the approximate probability of choosing product $a_j$ from offer set $S$ assuming that the sampled preference list is consistent with DAG $D$:

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

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 $\underline{f}(a_j, S, D) = \frac{\lambda(DuC(a_j, S))}{\lambda(D)}$ if $a_j \in h_D(S)$ and $0$, otherwise; $\overline{f}(a_j, S, D) = \frac{\lambda(DuC(a_j, S))}{\lambda(D)}$ if $a_j \in h_D(S)$, and $0$, otherwise; we have the following inequalities (see Corollary 1 in Section 3):

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

and for the exact and hard-to-compute probability of purchase $f(a_j, S, D)$,

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

The right column of Figure A5 illustrates that the percentage of transactions when the prediction of the item to be chosen is made 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., $\underline{f}(a_j, S, D)$, in only 4.04% of the instances for individuals without cycles in their preference graph, and in only 1.79% of the instances for individuals with cycles in their preference graph. In both cases, our prediction is the item with the highest probability of being purchased. This empirical observation provides good support for the use of $\hat{f}(a_j, S, D)$ as a proxy for the true and hard-to-compute probability of purchase $f(a_j, S, D)$.

In our reported results in Section 4 we use the following tractable formula to compute the posterior probabilities of purchase:

$$\tilde{f}(a_j, S, D) = \begin{cases} \frac{v_j \Psi_D(a_j)}{\sum_{a_k \in h_D(S)} v_k \Psi_D(a_k)}, & \text{if } a_j \in h_D(S), \\ 0, & \text{otherwise}. \end{cases}$$

This expression is intended to be a good approximation for the alternative approximation $\hat{f}(a_j, S, D)$, which we already know is a good approximation for the exact $f(a_j, S, D)$. We verify this in Figure A6. Therein, 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 Section 4.3). For all the panels in Figure A6 the average MAE (Mean Absolute Error) is below 0.5% which indicates that the posterior probability approximation $\tilde{f}(a_j, S, D)$ is very close to the posterior probability based on the lower bound of the DAG likelihood $f(a_j, S, D)$ in terms of predictive performance.
A6. Robustness check for the prediction results in Section 4

In this section we summarize the major empirical experiments conducted in order to check the robustness of the prediction results reported in Section 4.4. We start checking the robustness with respect to different data aggregation strategies, followed by the effect of accounting for no-purchase observations, the effect of different cutoff points between training and holdout sample data, and the effect of adding implicit candidate edges in Phase 2 of the DAG construction process.

A6.1. Robustness with respect to different data aggregation strategies

We start by showing the robustness of the results to changes in how we calibrate the benchmarks. In particular, in Section 4.4 we calibrated all the models separately on (i) customers who do not have cycles in their preference graph under the PO-MNL Promotion, and (ii) customers who have cycles in their preference graph under the PO-MNL Promotion model. Then we used previous separate calibrations but presented the joint prediction over all the individuals for each category.
of products, i.e., the weighted average prediction performance between both types of customers: with and without cycles in their preference graph. Here we show the prediction performance of our model versus the benchmarks using a display format similar to that in Section 4.4, but when calibrating both benchmarks on the set of all individuals. Note that there is a tension about the benchmarks here since one side they are estimated on a larger volume of data, but at the same time this extra volume comes at the expense of higher customer heterogeneity (i.e., individuals without and with cycles pool together for the estimation process).

Similarly to Figure 3, Figure A7 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. Note that in all the panels we calibrate the benchmarks on the set of all individuals and then separately make predictions for three subsets of customers: without and with cycles, and the entire population.

First, consider the left two panels in Figure A7. Here, we calibrate the PO-MNL Promotion and make predictions with all the models on the subset of individuals who do not have cycles in their preference graph. The “chi-square” score of PO-MNL Promotion model exhibits an average improvement of 9.77% over LC-MNL and 5.79% over RPL. Using the LC PO-MNL Promotion model, we can further boost performance, resulting in average improvement of 14.31% over LC-MNL and 10.28% over RPL.

Second, consider the middle column in Figure A7, where we calibrate the PO-MNL Promotion and make predictions with all the models on the subset of individuals that have cycles in their preference graph. We see that PO-MNL Promotion model exhibits an average improvement of 12.4% over LC-MNL and deterioration of 1.05% over RPL, while LC PO-MNL Promotion model leverages the performance to an average improvement of 13.47% over LC-MNL and 0.2% over RPL.

Third, consider the right panels in Figure A7, where we use the previous separate calibrations for all the models but report the joint prediction over all the individuals for each category of products. The performance here is a weighted average between the two types of customers: with and without cycles in the preference graphs, achieving significant improvements overall: PO-MNL Promotion model exhibits an average improvement of 14.32% over LC-MNL and 2.68% over RPL, while LC PO-MNL Promotion model shows an average improvement of 16.24% over LC-MNL and 4.74% over RPL.

Figure A8 presents scatterplots of the miss-rates, using a display format similar to that of Figure 4. From it, we observe that our model combinations obtain improvements of up to 8.92% under PO-MNL Promotion, and further improvements of up to 11.3% under LC PO-MNL Promotion over the benchmarks within the six panels.
Figure A7  Brand choice “chi-square” prediction results. Scatter plot of the average $\chi^2$ scores of all 27 product categories under the best of up to 10 LC-MNL [benchmark] and RPL [benchmark], vs. the PO-MNL Promotion [single class] and LC PO-MNL Promotion [multiclass]. Lower is better; the benchmark is outperformed for points above the 45° line. In all the panels we estimate the benchmarks on the set of all individuals. Left panels: we predict all the models only over individuals who do not have cycles in the preference graph under the PO-MNL Promotion model. Middle panels: we predict all the models only over individuals who have cycles in the preference graph under the PO-MNL Promotion model. Right panels: we combine both types of the predictions from left and middle columns such that we cover all the individuals.

The key observations can be summarized as follows: (i) The PO MNL Promotion model, especially the multiclass version of it, outperforms the state-of-the-art competitive benchmarks even when the latter ones are allowed to be estimated on a larger population of customers; (ii) the best results for our model are observed on the individuals without cycles in the first place, who are the ones displaying the most consistent behavior; and (iii) LC-MNL is more competitive than RPL with respect to $\chi^2$ scores, but is dominated by RPL in terms of miss rates. The last observation is different from what we report in Section 4.4, where RPL dominated LC-MNL with respect to both $\chi^2$ and miss rates. The larger volume of data to train the models seems to favor more LC-MNL than RPL.
Figure A8  Brand choice miss rate prediction results. Scatter plot of the average miss rate of all 27 product categories under the best of up to 10 LC-MNL [benchmark] and RPL [benchmark], vs. the PO-MNL Promotion [single class] and LC PO-MNL Promotion [multiclass]. Lower is better; the benchmark is outperformed for points above the 45° line. In all the panels we estimate the benchmarks on the set of all individuals. Left panels: we predict all the models only over individuals who do not have cycles in the preference graph under the PO-MNL Promotion model. Middle panels: we predict all the models only over individuals who have cycles in the preference graph under the PO-MNL Promotion model. Right panels: we combine both types of the predictions from left and middle columns such that we cover all the individuals.

A6.2. Robustness with respect to adding no-purchase observations

In order to streamline the comparison of the models in the empirical case study in Section 4.4 we did not include the no purchase observations in the prediction tasks since we did not have explicit data on the no-purchase alternatives in our dataset. Here, we demonstrate that the brand choice prediction results remain qualitatively the same when we include the no purchase option in our calibration and prediction tasks.

To this end, we approximately build the no-purchase observations from our data. In particular, we can approximately infer from the data the times the customer visits to the store to make at least one category purchase. Therefore, we can easily obtain instances when the customer visited the store but ended up not making a specific category purchase. However, these observations in the
data cannot be considered as the no-purchase instances since we do not know if the customer had the intent to make a category purchase and ended up choosing the no-purchase option. In fact, the number of store visits is around ten times higher than the number of purchases for some categories. As a result, in order to minimize the number of ”spurious” no-purchase observations inferred from the data, we first assume that the number of times a customer chooses the outside option is comparable to the number of times a customer makes a category purchase. In particular, we say that the number of no-purchases of every customer is equal to the number of times a customer buys her second most purchased product. It implies that the customers chose a no-purchase alternative on average \(\alpha T_c\) times, where \(T_c\) is the total number of times a customer made a category purchase and \(\alpha = 0.05\). We also show below that the obtained prediction results are robust to other values of \(\alpha\). As a result, we randomly sample the fixed portion of the no-purchase observations from the data on the store visits of the customers when they decided not to make a category purchase, and include these additional transactions into our dataset for every category. Then we use the same approach described in the main body of the paper to calibrate the models and test their predictive performance.

Analogously to Figure 3, Figure A9 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. First, consider the left two panels. Here, we calibrate the models on the subset of individuals that do not have cycles in their preference graph. The ”chi-square” score of PO-MNL Promotion model exhibits a moderate average deterioration of 2.79% over LC-MNL and an average improvement of 7.43% over RPL. Second, consider the middle column in Figure A9, where we calibrated the models on the subset of individuals that have cycles in their preference graph. We see that PO-MNL Promotion model exhibits an average improvement of 12.88% over LC-MNL and 4.17% over RPL.

In the left two panels in Figure A9, it is demonstrated that using LC PO-MNL Promotion model, we can further boost performance of the proposed methodology, resulting in average improvement of 14.64% over LC-MNL and 22.8% over RPL. Similarly, as illustrated in the middle column of Figure A9, LC PO-MNL Promotion model, capturing heterogeneity of customers to a greater extent, has an average improvement of 16.07% over LC-MNL and 7.43% over RPL.

Third, consider the right panels in Figure A9, where we use the previous separate calibrations but report the joint prediction over all the individuals for each category of products. The performance here is a weighted average between the two types of customers: with and without cycles in the preference graphs, achieving significant improvements overall: PO-MNL Promotion model exhibits an average improvement of 12.59% over LC-MNL and 3.48% over RPL, while LC PO-MNL Promotion model shows an average improvement of 16.18% over LC-MNL and 7.27% over RPL.
Figure A9  Brand choice prediction results with no purchase option included. We assume that the number of no-purchases of every customer is equal to the number of times a customer buys her second most purchased product. Scatter plot of the average $\chi^2$ scores of all 27 product categories under the best of up to 10 LC-MNL [benchmark] and RPL [benchmark], vs. the PO-MNL Promotion [single class] and LC PO-MNL Promotion [multiclass]. Lower is better; the benchmark is outperformed for points above the 45° line. Left panels: we estimate and predict only over individuals who do not have cycles in the preference graph under the PO-MNL Promotion model. Middle panels: we estimate and predict only over individuals who have cycles in the preference graph under the PO-MNL Promotion model. Right panels: we estimate each type of individuals separately and combine both types for the prediction.

Analogously to Figure 4, Figure A10 presents scatterplots of the miss-rates, using a display format similar to that of Figure A9. From it, we observe that we obtain improvements of between 1.28% and 10.04% under LC PO-MNL Promotion over the benchmarks in all six panels.

Like in the base case in Section 4.4 without the no-purchase option, we observe that the PO MNL Promotion model continues to outperform both benchmarks in most of the categories, with respect to $\chi^2$ and miss rate, especially for its multiclass version.

Next, we showcase the robustness of the predictive results for other values of $\alpha$. In particular, we test the predictive performance of the LC PO-MNL Promotion model versus the LC-MNL benchmark for $\alpha = 30\%$ and $\alpha = 40\%$ in Figures A11 and A12, respectively. Here we focus only on the LC-MNL benchmark provided its competitive performance and the heavy computational burden of RPL. We observe that the improvements of PO-MNL Promotion over the LC-MNL
Figure A10  Brand choice prediction results with no purchase option included. We assume that the number of no-purchases of every customer is equal to the number of times a customer buys her second most purchased product. Scatter plot of the average miss rate of all 27 product categories under the best of up to 10 LC-MNL [benchmark] and RPL [benchmark], vs. the PO-MNL Promotion [single class] and LC PO-MNL Promotion [multiclass]. Lower is better; the benchmark is outperformed for points above the 45° line. Left panels: we estimate and predict only over individuals who do not have cycles in the preference graph under the PO-MNL Promotion model. Middle panels: we estimate and predict only over individuals who have cycles in the preference graph under the PO-MNL Promotion model. Right panels: we estimate each type of individuals separately and combine both types for the prediction.

benchmark vary with $\alpha$ between 12.36% and 14.95% in terms of the $\chi^2$ score, and between 1.24% and 1.76% in terms of the miss rate.

Figures A11 and A12 confirm that the superior performance of PO MNL Promotion is robust to different fractions of no-purchases in the dataset.

A6.3. Robustness with respect to the split between test and training datasets

In all our experiments so far, the training set consists of the first 26 weeks and the test set consists of the last 26 weeks of transactions. Here we show the robustness of the results to changes in how we split the data into the training and test sets. We also include the no-purchase observations (with $\alpha = 20.5\%$).
Figure A11  Brand choice prediction results: with no purchase option included. Every customer is assumed to choose a no-purchase alternative on average $\alpha T_c$ times, where $T_c$ is the total number of times a customer made a category purchase and $\alpha = 30\%$. Scatter plot of the prediction scores of all 27 product categories under the best of up to 10 LC-MNL vs. LC PO-MNL Promotion [PO-MNL Promotion]. Lower is better; the benchmark is outperformed for points above the $45^\circ$ line. Left panel: miss rate. Right panel: $\chi^2$ score.

Figure A12  Brand choice prediction results: with no purchase option included. Every customer is assumed to choose a no-purchase alternative on average $\alpha T_c$ times, where $T_c$ is the total number of times a customer made a category purchase and $\alpha = 40\%$. Scatter plot of the prediction scores of all 27 product categories under the best of up to 10 LC-MNL vs. LC PO-MNL Promotion [PO-MNL Promotion]. Lower is better; the benchmark is outperformed for points above the $45^\circ$ line. Left panel: miss rate. Right panel: $\chi^2$ score.

When we reduce the volume of the training set to the first 21 weeks and enlarge the test set to the last 31 weeks of transactions, Figure A13 shows that the LC PO-MNL Promotion model
Figure A13 Brand choice prediction results: with no purchase option included. The training data consists of 21 weeks, and the test data consists of 31 weeks. Scatter plot of the prediction scores of all 27 product categories under the best of up to 10 LC-MNL [LC-MNL] vs. the LC PO-MNL Promotion [PO-MNL Promotion]. Lower is better; the benchmark is outperformed for points above the 45° line. Left panel: miss rate score. Right panel: \( \chi^2 \) score.

outperforms the LC-MNL benchmark by 16.14% and 1.02% based on the miss rate and \( \chi^2 \) scores, respectively.

When increasing the training set to the first 31 weeks and reducing the test set to the last 21 weeks of transactions, Figure A14 shows that the LC PO-MNL Promotion model outperforms the LC-MNL benchmark by 14.66% and 1.12% based on the miss rate and \( \chi^2 \) scores, respectively.

We observe a sustained relative performance of the PO MNL Promotion model over the LC-MNL model as we reduce or increase the training dataset.

A6.4. Robustness with respect to the addition of implicit candidate edges in the DAGs

Phase 2 in the DAG construction process is about the inclusion of implicit candidate edges in the DAG that identifies each individual. This is a heuristic step that assumes that the relative preference between two products is preserved regardless the promotion status of the products. That is, if from Phase 1 a full price version of \( a_j \) is preferred over the full price version of \( a_k \), then the promoted version \( a_{j+n} \) is also preferred to the promoted version \( a_{k+n} \). Similarly, if the Phase 1 preference is stated on the promoted versions, then the relative preference is extended to the corresponding full price versions. These Phase 2 edges have a low weight and some of them are the ones to be deleted in Phase 3 in case a cycle arises in the DAG. Of course, the heuristic could add spurious implicit candidate edges, and the final justification for the existence of the edges is their empirical performance.
Figure A14  Brand choice prediction results: with no purchase option included. The training data consists of 31 weeks, and the test data consists of 21 weeks. Scatter plot of the prediction scores of all 27 product categories under the best of up to 10 LC-MNL [LC-MNL] vs. the LC PO-MNL Promotion [PO-MNL Promotion]. Lower is better; the benchmark is outperformed for points above the 45° line. Left panel: miss rate score. Right panel: $\chi^2$ score.

To this end, Figure A15 illustrates that the LC PO-MNL Promotion model with implicit candidate edges outperforms the LC PO-MNL Promotion without implicit edges by 1.82% and 1.9% based on miss rate and $\chi^2$ scores, respectively. This observation provides enough support for their inclusion in the DAG. It allows us to conclude that adding implicit edges in the DAG construction process boosts the benefit in prediction performance because of the denser DAG which outweighs the biases from adding few spurious edges along the way.

A7. Comparison with the DAG-based behavioral models studied by Jagabathula and Vulcano (2018)

In this section we compare the predictive performance of the PO-MNL Promotion model with the PO-MNL Inertial and PO-MNL Censored models studied by Jagabathula and Vulcano (2018). Note that both PO-MNL Inertial and PO-MNL Censored models take into account the information about product promotions implicitly through modeling the consideration sets of customers via behavioral rules. Relying on the pre-specified behavioral assumptions this approach cannot consistently explain the purchasing behavior of all the customers. As a result, the DAGs of customers whose purchasing transactions are inconsistent with these assumptions are assumed to be empty (i.e., without edges), which reduce the representation of customer preferences to any standard random utility model such as the MNL or the LC-MNL. In particular, to run the prediction performance of PO-MNL Inertial and PO-MNL Censored for customers that have empty DAGs, we use the best of up to 10 LC-MNL [LC-MNL].
Recall that the approach taken in our paper for the DAG construction is different, since it is completely data-driven and accounts explicitly for promotion effects. The approach could still lead to cycles in the preference graph, which are then deleted in Phase 3 such that all customers are characterized by non-empty DAGs. Figure A16 illustrates the scatter plot of the $\chi^2$ scores of all 29 product categories under the LC PO-MNL Promotion vs. PO-MNL Inertial with clustering and PO-MNL Censored with clustering (see (Jagabathula and Vulcano 2018, Section 5)). In all the panels we calibrate the LC-MNL and LC PO-MNL Promotion models on the set of all individuals. First, consider the left two panels in Figure A16. Here, we calibrate the PO-MNL Inertial [top left panel] and PO-MNL Censored [bottom left panel] and represent the prediction performance of all the models over the subset of individuals that can be explained by behavioral assumptions. The $\chi^2$ score of LC PO-MNL Promotion model exhibits an average deterioration of 15.52% over PO-MNL Inertial with clustering and 18.68% over PO-MNL Censored with clustering. Second, consider the right column in Figure A16, where we calibrate the PO-MNL Inertial [top right panel] and PO-MNL Censored [bottom right panel] and represent the prediction performance of all the models over the subset of individuals that can not be explained by behavioral assumptions. In this case, both PO-MNL Inertial and PO-MNL Censored models are reduced to the LC-MNL model. We see that LC PO-MNL Promotion model exhibits an average improvement of 11.18% over PO-MNL Inertial and 2.14% over PO-MNL Censored.

Figure A17 presents scatterplots of the miss rates, using a display format similar to that of Figure A16. The insights are the same as in Figure A16. In the left column we calibrate the PO-MNL Inertial [top left panel] and PO-MNL Censored [bottom left panel] models and represent the
Figure A16  The layout is the same as Figure 3 in the paper by Jagabathula and Vulcano (2018). Scatter plot of the $\chi^2$ scores of all 29 product categories under the PO-MNL Inertial and PO-MNL Censored vs. the PO-MNL Promotion. In all the panels we calibrate the LC-MNL and LC PO-MNL Promotion models on the set of all individuals. Left panels: we estimate PO-MNL Inertial and PO-MNL Censored and make predictions for the models over the individuals whose purchasing transactions can be explained by the behavioral assumptions of PO-MNL Inertial [top left panel] or PO-MNL Censored [bottom left panel]. Right panels: Here, both PO-MNL Inertial and PO-MNL Censored models are reduced to LC-MNL model. We estimate PO-MNL Inertial and PO-MNL Censored and make predictions for the models over the individuals whose purchasing transactions can not be explained by the behavioral assumptions of PO-MNL Inertial [top right panel] and PO-MNL Censored [bottom right panel].

prediction performance of all the models over the subset of individuals who can be explained by behavioral assumptions. We observe that LC PO-MNL Promotion obtains an average deterioration of 1.36% over PO-MNL Inertial with clustering and of 3.66% over PO-MNL Censored with clustering. Then, in the right column in Figure A17, we calibrate the PO-MNL Inertial [top right panel] and PO-MNL Censored [bottom right panel] and represent the prediction performance of all the models over the subset of individuals that can not be explained by behavioral assumptions. In this case, both PO-MNL Inertial and PO-MNL Censored models are reduced to LC-MNL model. We notice that LC PO-MNL Promotion model exhibits an average improvement of 2.13% over PO-MNL Inertial and 2.53% over PO-MNL Censored.

Even though the results in Figures A16 and A17 show an average dominance of the behavioral models over the PO MNL Promotion optimization model with respecto to both $\chi^2$ and miss rates, the presence of points above the diagonal indicates that for some categories PO MNL Promotion
Figure A17  The layout is the same as Figure 5 in the paper by Jagabathula and Vulcano (2018). Scatter plot of the miss rate scores of all 29 product categories under the PO-MNL Inertial and PO-MNL Censored vs. the PO-MNL Promotion. In all the panels we calibrate the LC-MNL and LC PO-MNL Promotion models on the set of all individuals. Left panels: we estimate PO-MNL Inertial and PO-MNL Censored and make predictions for the models over the individuals whose purchasing transactions can be explained by the behavioral assumptions of PO-MNL Inertial [top left panel] or PO-MNL Censored [bottom left panel]. Right panels: Here, both PO-MNL Inertial and PO-MNL Censored models are reduced to LC-MNL model. We estimate PO-MNL Inertial and PO-MNL Censored and make predictions for the models over the individuals whose purchasing transactions can not be explained by the behavioral assumptions of PO-MNL Inertial [top right panel] and PO-MNL Censored [bottom right panel].

still dominates. In order to characterize those categories, in Figure A18 we report the loyalty score of each category computed on the training data (left panel). Then, in the middle and right panels we explore possible correlations between the percentage of $\chi^2$ improvement of the behavioral models with respect to PO MNL Promotion (vertical axis) vs. loyalty score (horizontal axis). We note a negative correlation for PO MNL Promotion improvements with respect to both PO MNL Inertial and Censored models, meaning that the behavior of customers for most of the categories with low loyalty index (which exhibit the least stickiness in customers’ preferences) are better explained by the PO MNL Promotion model, as it is the case for customers represented by empty DAGs in the Jagabathula and Vulcano’s approach.

These findings suggest that the practitioners might use the PO-MNL Inertial and Censored models for categories with high loyalty index, and within them, for customers having non-empty
Figure A18  Loyalty scores and improvements of PO MNL Promotion over behavioral DAG-based models. The layout is similar to Figure 4 in Jagabathula and Vulcano (2018). Left panel: Loyalty scores across categories, sorted in descending order. Middle panel: Scatter plot and linear regression of the percentage $\chi^2$ improvement of PO MNL Promotion with respect to PO MNL Inertial (vertical axis) vs. loyalty score. Middle panel: Scatter plot and linear regression of the percentage of $\chi^2$ improvement of PO MNL Promotion with respect to PO MNL Inertial (vertical axis) vs. loyalty score (horizontal axis). Right panel: Scatter plot and linear regression of the percentage of $\chi^2$ improvement of PO MNL Promotion with respect to PO MNL Censored (vertical axis) vs. loyalty score (horizontal axis).

DAGs. Other than this, the use of the PO-MNL Promotion model proposed in this paper leads to more effective predictions.

A8. Optimization of personalized promotions

We now show that the set of constraints (16)–(19) ensures that $p$’s are normalized attraction values. Recall that $z_j = 1$ for all products $a_j$ in the set $h_D(S(y))$ of heads in the subgraph of the transitive closure of $D$ restricted to the set $S(y)$.

**Lemma A3.** Suppose $(0 \leq p_j \leq 1; a_j \in S_A)$ satisfy (16)–(19), then

$$p_j = \begin{cases} 0, & \text{if } z_j = 0, \\ \frac{v_{\Psi_D(a_j)}}{1 + \sum_{k: a_k \in S_A} v_{\Psi_D(a_k)}}, & \text{if } z_j = 1. \end{cases}$$

**Proof:** To simplify notation, let $w_j$ denote $v_{\Psi_D(a_j)}$ for each $a_j \in S_A$. For convenience, we reproduce the set of constraints (16)–(19) below:

$$p_j \leq z_j \quad \forall \ a_j \in S_A,$$  \hspace{1cm} (A7)
\[
p_0 + \sum_{j: a_j \in S_A} p_j = 1, \tag{A8}
\]
\[
0 \leq p_j \leq w_j p_0, \quad \forall \ a_j \in S_A \tag{A9}
\]
\[
p_0 + z_j - 1 \leq p_j / w_j \quad \forall \ a_j \in S_A. \tag{A10}
\]

Let \( S \) denote the set \( \{ a_j : z_j = 1 \} \), consisting of all the product indices such that \( z_j = 1 \). It immediately follows from (A7) that \( p_j = 0 \) for all \( a_j \in S_A \setminus S \) (since \( z_j = 0 \) therein). Then, for all \( a_j \in S_A \setminus S \), (A9) trivially holds and (A10) reduces to \( p_0 \leq 1 \) (which also trivially holds).

Now, for any \( j \) such that \( a_j \in S \) (and hence, \( z_j = 1 \)), we have from (A9) and (A10) that

\[
0 \leq p_j \leq w_j p_0 \quad \text{and} \quad p_0 \leq p_j / w_j.
\]

It thus follows that \( p_j = p_0 w_j \) for all \( a_j \in S \). We now obtain from (A8) that

\[
p_0 + \sum_{a_j \in S_A} p_j = 1 \implies p_0 + \sum_{a_j \in S} p_j = 1 \implies p_0 + \sum_{a_j \in S} p_0 w_j = 1 \implies p_0 = 1 / \left( 1 + \sum_{a_j \in S} w_j \right),
\]

where the first implication follows because \( p_j = 0 \) for all \( a_j \in S_A \setminus S \) and the second implication follows because \( p_j = p_0 w_j \) for all \( a_j \in S \). Since \( w_j > 0 \) for all \( a_j \), it follows that \( p_0 \leq 1 \), as needed.

We have thus obtained that

\[
p_j = 0 \text{ for all } a_j \in S_A \setminus S \text{ and } p_j = w_j p_0 = w_j / \left( 1 + \sum_{a_k \in S} w_k \right) \text{ for all } a_j \in S.
\]

In other words, we have that \( p_j = w_j z_j / \left( 1 + \sum_{a_k \in S_A} w_k z_k \right) \), which follows from the definition of \( S \).

The result of the lemma now holds. \( \square \)

### A9. Robustness check for the promotion optimization

In the revenue results in Section 5 our prediction assumed that when arriving at the store every customer would buy the product with highest probability of being purchased under PO-MNL Promotion model. An alternative objective would be to compute expected revenues accounting for the probabilities of purchase of every single product on offer. Figure A19 illustrates the results of running personalized promotions under the modified objective function of the optimization problem. Since the new promotion optimization results, illustrated in Figure A19, almost exactly resemble the ones in Figure 5, we conclude that all the insights remain qualitatively the same under the updated formulation of the promotion optimization problem.
A10. Managerial insights: Factors affecting improvements from personalization of promotions

The revenue improvements from personalization vary across different customers in each category. To explain this variation, for each category, we consider three different customer-level characteristics:

1. **Brand loyalty**, measured as the percentage of store visits a customer buys her most frequently purchased brand from the category;

2. **Purchase frequency**, measured as the number of purchases a customer makes from the category over the training data (26 weeks); and

3. **Promotion sensitivity**, measured as the percentage of store visits a customer buys a promoted product from the category.

We regressed the revenue improvement for each customer and category combination against the brand loyalty, purchase frequency and promotion sensitivity variables. Table A1 reports the results from fitting four different models:

\[
\text{RevImpr}_{i,c} = \beta_{01} \cdot \text{Cat}_c + \beta_{11} \cdot \text{Bloyalty}_{i,c} + \varepsilon_{i,c} \quad \text{(Model 1)}
\]

\[
\text{RevImpr}_{i,c} = \beta_{02} \cdot \text{Cat}_c + \beta_{22} \cdot \text{PurFreq}_{i,c} + \varepsilon_{i,c} \quad \text{(Model 2)}
\]

\[
\text{RevImpr}_{i,c} = \beta_{03} \cdot \text{Cat}_c + \beta_{33} \cdot \text{PromSens}_{i,c} + \varepsilon_{i,c} \quad \text{(Model 3)}
\]

\[
\text{RevImpr}_{i,c} = \beta_{04} \cdot \text{Cat}_c + \beta_{14} \cdot \text{Bloyalty}_{i,c} + \beta_{24} \cdot \text{PurFreq} + \beta_{34} \cdot \text{PromSens}_{i,c} + \varepsilon_{i,c} \quad \text{(Model 4)}
\]
where the variables Bloyalty\(_{i,c}\), PurFreq\(_{i,c}\), and PromSens\(_{i,c}\) respectively denote the brand loyalty, purchase frequency, and promotion sensitivity computed for customer \(i\) under category \(c\). The brand loyalty and purchase frequency variables were computed using the training data. To ensure exogeneity, the promotion sensitivity variable was computed using the data from the previous year (2006). The variable Cat\(_c\) is an indicator variable denoting category \(c\) to capture category fixed effects. Finally, RevImpr\(_{i,c}\) is the average revenue improvement from personalization for customer \(i\) under category \(c\), computed over the holdout sample.

The results from the regressions are consistent and intuitive. The benefits from personalization are negatively correlated with brand loyalty (Model 1) and purchase frequency (Model 2) but positively correlated with promotion sensitivity (Model 3). In other words, customers who purchase infrequently and concentrate their purchases only on a few brands are harder to persuade to switch to more profitable brands through personalized promotions. On the other hand, customers who frequently purchase promoted items are easier to be influenced by personalizing promotions. These findings are consistent in a multiple regression of the revenue improvement against all three variables (Model 4). The coefficients in the multiple regression are all statistically significant, indicating that all three factors together influence the brand switching behavior of customers in response to promotions.

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Revenue Improvement (%)</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brand loyalty</td>
<td>-14.415(***)</td>
<td></td>
<td>-20.514(***)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-2.742)</td>
<td></td>
<td>(-3.521)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purch. frequency</td>
<td></td>
<td>-1.108(***)</td>
<td>-1.218(***)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-3.133)</td>
<td>(-3.224)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prom. sensitivity</td>
<td></td>
<td>14.575(***)</td>
<td>13.891(***)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3.666)</td>
<td>(3.610)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Category FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>No. Observations</td>
<td>57,059</td>
<td>57,059</td>
<td>57,059</td>
<td>57,059</td>
<td></td>
</tr>
<tr>
<td>R-squared:</td>
<td>0.001</td>
<td>0.003</td>
<td>0.001</td>
<td>0.006</td>
<td></td>
</tr>
</tbody>
</table>

\(t\) statistics in parentheses

\* \(p<0.1\), \** \(p<0.05\), \*** \(p<0.01\)

Table A1 Individual-level regressions with product category fixed effects.

A11. Additional insights on brand loyalty and the number of promoted items

In this section, we provide additional descriptive statistics to gain insights into the extent of brand loyalty of the customers and the number of promoted items in our dataset. Table A2 presents these statistics.
The second to sixth columns in the table report the statistics describing the distribution of the number of unique brands purchased by the customers in each category; specifically, the columns report the mean, the standard deviation, and the first, second, and the third quartiles of the distribution, respectively. We note that on average customers purchase no more than 4 unique brands in the training data, indicating that customers have strong preferences. This observation complements the loyalty scores reported in Figure A18 and explains the significant performance gains that our model obtains over the benchmark models as reported in Figure 3.

Columns seven and eight report the average number of products our method offers on promotion across all the transactions in the holdout sample. Column seven reports this number when existing mass promotions are ignored and column eight reports the number of promoted products on top of the products that are already on mass promotion. We note that most of these numbers are less than 1, indicating that at optimality, our method offers only a small number of products on promotion. This observation explains why our method is able to extract most of the revenues even with the constraint of offering at most one product on promotion; see Section 5.3.

In order to provide a partial explanation for the small number of products that are put on promotion, the last two columns of the table report the average offer set size and the average number of products that can be potentially promoted (NonDom) across all the transactions. To calculate the number of products that can be potentially promoted, consider a transaction in which the offer set is \( S \) and the customer has DAG \( D \). We note that product \( a_j \) will not be promoted if there is another product \( a_i \in S \) such that the non-promoted copy of \( a_i \) (and consequently, the promoted copy of \( a_i \)) is preferred over the promoted copy of \( a_j \) in DAG \( D \). The reason is that product \( a_j \) will not be purchased whether promoted or not because either the promoted or the non-promoted copy of \( a_i \) will be offered to the customer. We call such a product \( a_j \) a dominated product and any product that is not dominated, a non-dominated product. Given this, NonDom reports the average number of non-dominated products across all the transactions in the holdout sample.

We observe from the table that customer DAGs are such that the average number of non-dominated products is far smaller than the average offer set size. Because the number of non-dominated products is an upper bound on the number of products that will promoted, this table provides a partial explanation as to why at optimality, only a small number of products are promoted.
<table>
<thead>
<tr>
<th>Category</th>
<th># Unique brands purchased</th>
<th># prom. items</th>
<th>AvOS</th>
<th>NonDom</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>stdev</td>
<td>1st quart.</td>
<td>2nd quart.</td>
</tr>
<tr>
<td>beer</td>
<td>2.09</td>
<td>1.29</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>carbbev</td>
<td>2.91</td>
<td>1.31</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>cigets</td>
<td>1.29</td>
<td>0.52</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>coffee</td>
<td>2.17</td>
<td>1.19</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>colder</td>
<td>3.03</td>
<td>1.37</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>deod</td>
<td>2.03</td>
<td>1.04</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>factiss</td>
<td>1.90</td>
<td>0.87</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>fzdigest</td>
<td>3.36</td>
<td>2.10</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>fzpizza</td>
<td>2.52</td>
<td>1.33</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>hhclean</td>
<td>2.90</td>
<td>1.40</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>hotdog</td>
<td>2.09</td>
<td>1.04</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>laundet</td>
<td>1.96</td>
<td>1.07</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>margbutr</td>
<td>1.93</td>
<td>1.02</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>mayo</td>
<td>1.36</td>
<td>0.55</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>milk</td>
<td>2.36</td>
<td>1.17</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>mustketc</td>
<td>2.29</td>
<td>0.92</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>paptowl</td>
<td>2.15</td>
<td>1.08</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>peanbutr</td>
<td>1.65</td>
<td>0.75</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>saltsnck</td>
<td>3.67</td>
<td>1.89</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>shamp</td>
<td>2.15</td>
<td>1.14</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>soup</td>
<td>2.94</td>
<td>1.46</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>spagsauc</td>
<td>2.19</td>
<td>1.25</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>sugarsub</td>
<td>1.23</td>
<td>0.47</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>toitisu</td>
<td>2.01</td>
<td>1.06</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>toothbr</td>
<td>2.06</td>
<td>1.00</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>toothpa</td>
<td>1.65</td>
<td>0.74</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>yogurt</td>
<td>2.41</td>
<td>1.30</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

**Table A2** Relevant summary statistics from the data. The column '# Unique brands purchased' reports the distribution of the number of unique brands bought by a customer over the training period; specifically, the columns mean, stdev, 1st quart., 2nd quart., and 3rd quart. report the mean, the standard deviation, the 1st, 2nd, and 3rd quartiles, respectively. The column 'w./o. mass' reports the average number of products offered on promotion by our method, across all the transactions in the holdout sample when existing mass promotions are ignored. Similarly, the column 'w. mass' reports the average number of products offered on promotion by our method across all the transactions in the holdout sample on top of the products that are already on mass promotion. The columns 'AvOS' and 'NonDom' report the average number of vendors and the average number of non-dominated vendors in the offer set, respectively, across all the transactions in the holdout sample; see text for the definition of non-dominated vendors.
A12. Promotion optimization: Mitigation of the reference price effect

In this section, we illustrate how to control the negative long-term impact of the reference price effect within our promotion optimization framework. As it was mentioned in Section 6, repeated price promotions might reduce the reference price of the involved brands. Consequently, it might negatively affect the future customer reactions to price promotions and hence, the long-term revenue for the retailer.

To illustrate one possible way to control for this effect, we define the promotion repetition index in the following way

\[
\text{PrR}_\Delta = \frac{1}{|U|} \sum_{i \in U} \sum_{t=1}^{T_i} \mathbb{I}[\Delta_{it}] \cdot \frac{T_i}{|U|} \cdot 100\%,
\]

where \( U \) denotes the set of users, \( T_i \) is the number of transactions in the testing dataset for customer \( i \), and \( \mathbb{I}[\Delta_{it}] \) is the indicator function that takes the value 1 if the brand that is promoted at time \( t \) for individual \( i \) was also promoted at least once over the last \( \Delta \) weeks, and zero otherwise.

We compute this metric across the 27 product categories, accounting only for individuals who made at least two purchases over those 26 weeks. This index, which varies from 0% to 100%, quantifies the potential for lowering the reference price of the brands because of the repeated price promotions offered to customers, where higher indices indicate more frequent promotion repetition.

In Table A3, we compute the promotion repetition index in the test dataset after adding the constraint of promoting at most one item (see Section 5.3) for \( \Delta = 2, 3, \) and 4 weeks in the second, third, and fourth columns, respectively. Naturally, the promotion repetition index increases with \( \Delta \).

In order to mitigate the reference price effect in our promotion optimization framework in Section 5, we incorporate an additional constraint to the promotion optimization problem (20) to ensure that we do not repeatedly promote the same brand over \( \Delta \) consecutive weeks. Note that enforcing this constraint in the optimization problem (20) will make promotion repetition index equal to 0%. Figure A20 illustrates the results of running single item personalized promotions with this additional constraint incorporated to the optimization problem. We observe revenue improvements of 16.1%, 14.16%, and 12.47% when \( \Delta = 2, 3 \) and 4, respectively. Recall that the promotion optimization revenue improvement (capping the number of items to promote at one) when ignoring the promotion repetition constraint is 23.88%, see Section 5.3. In this regard, by explicitly avoiding the frequent promotion of the same item, the retailer sacrifices some short-term revenues in order to mitigate the reference price effect in the long-term revenues. Yet, the revenue improvement remains significant when compared to the baseline (current) situation.
Jagabathula, Mitrofanov and Vulcano: *Personalized Retail Promotions*

**Table A3**  Promotion repetition index for $\Delta = 2$, 3, and 4 weeks.

<table>
<thead>
<tr>
<th>product</th>
<th>Promotion repetition</th>
<th>$\Delta=2$</th>
<th>$\Delta=3$</th>
<th>$\Delta=4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>beer</td>
<td></td>
<td>27.53</td>
<td>33.43</td>
<td>37.69</td>
</tr>
<tr>
<td>carbbev</td>
<td></td>
<td>39.75</td>
<td>45.11</td>
<td>48.44</td>
</tr>
<tr>
<td>cigets</td>
<td></td>
<td>33.00</td>
<td>39.06</td>
<td>43.29</td>
</tr>
<tr>
<td>coffee</td>
<td></td>
<td>13.64</td>
<td>17.56</td>
<td>21.02</td>
</tr>
<tr>
<td>coldcer</td>
<td></td>
<td>27.95</td>
<td>33.60</td>
<td>37.21</td>
</tr>
<tr>
<td>deod</td>
<td></td>
<td>14.03</td>
<td>18.32</td>
<td>21.84</td>
</tr>
<tr>
<td>factiss</td>
<td></td>
<td>17.00</td>
<td>23.05</td>
<td>28.10</td>
</tr>
<tr>
<td>fzdinent</td>
<td></td>
<td>19.41</td>
<td>23.36</td>
<td>26.40</td>
</tr>
<tr>
<td>fpizza</td>
<td></td>
<td>15.13</td>
<td>19.21</td>
<td>22.92</td>
</tr>
<tr>
<td>hhclean</td>
<td></td>
<td>21.01</td>
<td>26.20</td>
<td>30.74</td>
</tr>
<tr>
<td>hotdog</td>
<td></td>
<td>8.18</td>
<td>11.18</td>
<td>14.18</td>
</tr>
<tr>
<td>laundet</td>
<td></td>
<td>11.02</td>
<td>15.45</td>
<td>19.46</td>
</tr>
<tr>
<td>margbutr</td>
<td></td>
<td>18.26</td>
<td>24.15</td>
<td>28.47</td>
</tr>
<tr>
<td>mayo</td>
<td></td>
<td>9.52</td>
<td>13.83</td>
<td>17.24</td>
</tr>
<tr>
<td>milk</td>
<td></td>
<td>13.38</td>
<td>15.34</td>
<td>16.57</td>
</tr>
<tr>
<td>mustkete</td>
<td></td>
<td>10.28</td>
<td>13.55</td>
<td>15.98</td>
</tr>
<tr>
<td>paptowl</td>
<td></td>
<td>25.65</td>
<td>33.50</td>
<td>39.21</td>
</tr>
<tr>
<td>peanutr</td>
<td></td>
<td>4.33</td>
<td>6.20</td>
<td>7.69</td>
</tr>
<tr>
<td>saltsnck</td>
<td></td>
<td>33.91</td>
<td>39.02</td>
<td>42.23</td>
</tr>
<tr>
<td>champ</td>
<td></td>
<td>16.49</td>
<td>21.39</td>
<td>26.40</td>
</tr>
<tr>
<td>soup</td>
<td></td>
<td>40.30</td>
<td>47.73</td>
<td>52.67</td>
</tr>
<tr>
<td>spagsauc</td>
<td></td>
<td>17.93</td>
<td>23.17</td>
<td>27.11</td>
</tr>
<tr>
<td>sugarsub</td>
<td></td>
<td>14.55</td>
<td>21.52</td>
<td>27.86</td>
</tr>
<tr>
<td>toitisu</td>
<td></td>
<td>26.36</td>
<td>34.79</td>
<td>41.65</td>
</tr>
<tr>
<td>toothbr</td>
<td></td>
<td>12.51</td>
<td>15.99</td>
<td>18.84</td>
</tr>
<tr>
<td>toothpa</td>
<td></td>
<td>14.15</td>
<td>18.79</td>
<td>23.81</td>
</tr>
<tr>
<td>yogurt</td>
<td></td>
<td>18.00</td>
<td>22.26</td>
<td>24.89</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td><strong>19.38</strong></td>
<td><strong>24.32</strong></td>
<td><strong>28.24</strong></td>
</tr>
</tbody>
</table>

**Figure A20**  Revenue performance of promoting a single item. Scatter plot of the expected revenue per customer transaction after single item promotion optimization when we do not repeatedly promote the same brand over $\Delta$ weeks [Prom. opt. revenue] vs. the realized revenue per customer transaction obtained from the data [Realized revenue] across 27 product categories, for all individuals based on PO-MNL Promotion model. The baseline is outperformed for points above the 45-degree line.