Quantitative Marketing and Economics
https://doi.org/10.1007/s11129-024-09287-y

The customer journey as a source of information

Nicolas Padilla[1] · **Eva Ascarza**[2] · **Oded Netzer**[3]

Received: 6 November 2023 / Accepted: 22 August 2024

Abstract
We introduce a probabilistic machine learning model that fuses customer click-stream data and purchase data within and across journeys. This approach addresses the critical business need for leveraging first-party data (1PD), particularly in environments with infrequent purchases, which are characterized by minimal or no prior purchase history. Combining data across journeys poses a challenge because customers' needs might vary across different purchase occasions. Our model accounts for this "context heterogeneity" using a Bayesian non-parametric Pitman-Yor process. By drawing from within-journey, past journeys, and cross-customer behaviors, our model addresses the "cold start problem," enabling firms to predict customer preferences even without prior interactions. Notably, the model continuously updates the inferred preferences as customers interact with the firm. We apply this model to data from an online travel platform, revealing significant benefits from consolidating 1PD from both current and previous customer journeys. This integration enhances managers' understanding of customer needs, allowing for more effective personalization of marketing tactics, such as retargeting efforts and product recommendations, to better align with customers' dynamic preferences.

Keywords Customer journey · Probabilistic machine learning · Bayesian nonparametrics · First-party data · Cold start · Click-stream data · Customer search

The authors thank Asim Ansari, Bruce Hardie, Anja Lambrecht, Vicki Morwitz, and Eric Schwartz for their comments. In writing the manuscript, the authors used ChatGPT (version 4) to correct grammatical mistakes and to refine (prompted) sentences or paragraphs, but never to create new content.

✉ Nicolas Padilla
npadilla@london.edu

Eva Ascarza
eascarza@hbs.edu

Oded Netzer
onetzer@gsb.columbia.edu

[1] London Business School, London NW1 4SA, UK

[2] Harvard Business School, Harvard University, Cambridge, MA 02138, USA

[3] Columbia Business School, Columbia University, New York, NY 10027, USA

Published online: 05 November 2024

JEL Classification M31 · C11

1 Introduction

Amid stricter privacy regulations on third-party sources like cookies and data brokers, marketers are increasingly leveraging first-party data (1PD) for marketing strategies, focusing on insights from their internal channels (Murphy, 2022). The implications of limited data access are more pronounced in situations with infrequent purchases. To compensate for the limited history of transactions from the same customer, firms often rely on digital footprints from customer-firm interactions on the company's own channels, such as search queries, clicks, and filtering of alternatives. This collection of behaviors is what we refer to as the *first-party customer journey*, with each journey encapsulating the entirety of behaviors collected by the company as customers navigate towards fulfilling distinct needs. Despite the potential of 1PD, many firms struggle to fully leverage these data sources (Huberman, 2021), particularly when it comes to combining various types of customer interactions and weighting them appropriately across different journeys.

This research proposes a modeling approach that enables firms to harness relevant information from first-party customer journeys by integrating data from various customer interactions with the company. Our method utilizes information across different journeys and customers, aiming to enhance understanding of customer needs, even when historical data is scarce. The primary challenge lies in extracting valuable insights from pre-purchase interactions, as these interactions do not always result in a purchase. Additionally, determining the appropriate weighting and integration of pre-purchase data with actual purchase information poses a significant challenge. Moreover, the variability in customer needs across different journeys complicates the ability to leverage data across journeys.

To address these issues, we introduce a probabilistic machine learning approach that links customer click-stream data over the course of a journey and integrates that information across journeys, and across the customer's history of purchases. The model accounts for what we call *context heterogeneity*, which are journey-specific preferences that capture customers' journey-specific needs and allow us to combine information across journeys with possibly varying needs. We model the journey decisions on what to search, what to click, and what to buy to be both a function of customers' stable preferences and the unique needs of the focal journey's context. Intuitively, contexts are unobserved states that capture need-specific preferences (e.g., food preferences in a special dinner might differ from preferences in a quick lunch). We uncover those contexts non-parametrically using a Pitman-Yor process, which allows for the creation of new contexts that have not been previously observed as new journey observations arrive.

The model leverages three primary sources of information collected from the company's own channels: (1) within-journey behavior, such as search queries and clicks during the focal journey; (2) past journeys' behavior, encompassing previous searches, clicks and purchases; and (3) cross-customer behavior, analyzing actions of other customers with similar search patterns. The within-journey data helps identify unique,

journey-specific preferences, while data from past journeys reveals the customer's stable preferences. Cross-customer information supplements potentially sparse data from the first two sources, enhancing our model's ability to predict preferences. This is particularly useful in cases where customers are unidentifiable — such as prior to logging in or their first purchase — and when past behaviors cannot be linked to a specific customer, for instance, due to disabled cookies or privacy regulations. By leveraging within journey information and integrating across-customer data from similar journeys, our approach effectively addresses the 'cold start problem,' enabling firms to infer preferences even before customers make their first purchase or are otherwise identified.

The model offers valuable insights for managers, highlighting the importance of information collected during the customer journey for a near-real-time analysis of customer preferences as the customer journey progresses. Dynamically assessing consumers' preferences, allows marketers to improve the personalization of retargeting efforts and product recommendations. Additionally, the contexts revealed by the model provide useful dynamic segmentation dimensions aiding in the design of user experiences or product features tailored to specific customer needs.

Our modeling approach is particularly suited for dynamically inferring customer preferences at the journey level, enhancing the effectiveness of personalized marketing strategies. Compared to structural search models, our method is more flexible, does not make assumptions about click order in the search process, reduces computational complexity, scales to large datasets, and integrates diverse data types, including past journeys, search queries, and within journey click-stream data. Compared to traditional machine learning (ML) approaches, our model maintains predictive validity and is more robust across prediction tasks, striking a balance between flexibility and interpretability. Our model permits the identification of interpretable latent journey-specific customer needs and a combination of both observed and unobserved heterogeneity.

We apply the model to customer journey data from a major online travel platform, allowing the firm to dynamically infer customer preferences for air travel attributes such as airline alliances, number of stops, and price sensitivity. For example, relative to what the platform can predict at the start of the customer journey (the query page), incorporating the initial two clicks from the current journey improves the prediction of a customer's preferred airline alliance by 25%, with accuracy increasing by 73% after five clicks. This demonstrates the model's ability to refine predictions in real-time as the journey progresses, a sharp contrast with traditional panel-data models (e.g., Rossi et al., 1996), which rely solely on purchase data and cannot update inferences based on real-time 1PD before a purchase is made. For example, our model predicts the actual product chosen ten times more accurately than models relying solely on historical purchase data. In a retargeting exercise, our model enhances business outcomes, such as increasing the click-through rate (CTR) for recommended products by as much as 28% relative to a heuristic retargeting using the last clicked product.

Our model addresses the "cold start problem," which arises when there is no historical data available for customers. We mitigate this in two ways. First, our model integrates information from other customers' journeys through their search queries. For example, our model improves the prediction of airline alliance choices by 38% at the start of a new journey compared to models that do not incorporate past journey

data. Second, our model leverages data from the current journey itself to compensate for this gap. With just five clicks in the current journey, our model achieves predictive power similar to models that utilize past journey data. This integration of data across first-party customer journeys is particularly valuable for understanding the preferences of new or unidentifiable customers—a growing challenge given the increasing prevalence of platforms that allow search without logging in or that are restricted in storing customer historical data.

Beyond enhancing predictions, our model provides valuable insights into the different types of customer journeys by uncovering 22 distinct contexts, each reflecting specific customer needs and preferences. For instance, one context identified by the model involves "one-way solo domestic trips," characterized by last-minute flight searches, unaccompanied passengers, and a preference for major airlines. Another context, termed "Family vacations in the Caribbean," includes customers seeking roundtrip flights for group travel, with a strong preference for non-stop routes and bookings made several months in advance.

Main contributions and related literature

Overall, this paper contributes to the literature by introducing a method to infer customer preferences in scenarios characterized by sparse purchase history and varying customer needs between transactions. This research is crucial for data-driven organizations aiming to optimize their use of 1PD amid tightening data regulations and increasing emphasis on customer privacy.

Our research builds on the rich literature in marketing that uses transactional data to estimate consumer preferences at the individual level (Rossi et al., 1996; Allenby & Rossi, 1998). These models are widely used in contexts where individual transactions are available, but challenges arise in situations with limited transaction data. The digital environment offers opportunities to observe the customer journey, including click-stream activity and past purchases, which we seek to combine effectively in our model.

Methodologically, our work aligns with probabilistic machine learning models in marketing (Dew et al., 2024), particularly Bayesian non-parametric models (Ansari & Mela, 2003; Kim et al., 2004; Braun & Bonfrer, 2011; Bruce, 2019; Dew et al., 2020; Boughanmi & Ansari, 2021). We extend this literature by proposing a Pitman-Yor process to infer the distribution of contexts, essential for fusing information across different customer journeys. This approach contrasts with other models of heterogeneous purchase occasions like Latent Dirichlet Allocation (LDA) (e.g., Jacobs et al., 2016), allowing us to non-parametrically determine the number of contexts from the data and separate how past journeys relate to a focal journey. Similar to Liu and Toubia (2018), we link search queries to customer preferences; however, we extend this work by leveraging correlations across multiple behaviors (e.g., queries, clicks, purchases) and product attributes, making it applicable to sparse, highly specific purchase data (e.g., flights). Additionally, our model accounts for customer-specific heterogeneity, allowing the use of past journeys in inferring preferences. Our research also relates to the literature on context-dependent product recommendations (e.g., Sarwar et al., 2001, Hidasi et al., 2016, Yoganarasimhan, 2020). While existing methods often require

extensive data on customer interactions with specific products, our model extracts preferences for attributes, making it applicable in settings with large, evolving product assortments.

Our work also contributes to the vast and rich literature on consumer search and the use of click-stream data (e.g., Montgomery et al. 2004; Kim et al. 2010; Seiler 2013; Donnelly et al. 2024; see Honka et al. 2024 for an extensive review) that uses within-journey information to infer customer preferences and predict purchase behavior. These studies have shown that browsing patterns and click-stream data can provide valuable insights into future customer actions. However, most of these approaches focus on a single journey, often overlooking the insights that previous journeys can offer (a notable exception is Morozov et al., 2021, who leverage a panel data of consumer searches). This limits their ability to capture the full scope of customer heterogeneity, both within and across journeys. Our research advances this field by incorporating not only click-stream data but also search queries, providing a more comprehensive analysis of customer behavior. By integrating data from multiple journeys, our approach captures the variability in preferences that can exist from one journey to another, addressing gaps in the existing literature. This integration allows for a deeper understanding of how customers' searching and clicking behaviors change from one purchase occasion to another, ultimately leading to a better understanding of consumer decisions.

Substantively, our work contributes to the literature on leveraging alternative sources of information when data on the main behavior of interest is limited (Iyengar et al., 2003; Feit et al., 2013; Liu & Toubia, 2018; Padilla & Ascarza, 2021). We highlight the value of extracting information from both current and previous customer journeys to infer preferences in the current journey. In terms of managerial implications, our findings contribute to the literature on the practical value of customer behavior data. Notable works by Rossi et al. (1996) and Smith et al. (2023) have shown that using historical purchase data can significantly improve the effectiveness of marketing strategies, from enhancing direct marketing outcomes to boosting profit margins through personalized pricing. Our study emphasizes the critical role of real-time customer data, including within-journey click-stream data, in refining personalization tactics and augmenting marketing efforts.

The paper continues as follows. Section 2 details our modeling framework and compares it with alternative approaches. Section 3 describes the empirical setting and estimation procedure. Section 4 presents the results, and Section 5 assesses the value of 1PD. We conclude with a discussion on the model's generalizability, potential limitations, and future research directions.

2 Model

2.1 Model overview

We propose a modeling framework that allows firms to extract the information contained in the various first-party interactions customers have with the focal firm. To that end, we define the *first-party customer journey* as the set of all the interactions

observed by a firm during the course of the customer journey aimed at satisfying the customer journey-specific needs. We highlight several components of this definition. First, a first-party customer journey only includes the interactions that the firm collects on its own platforms (e.g., web, apps, call center). We exclude interactions with competing firms aimed to satisfy the same need, as this information is often unobserved to the focal firm. Note that the customer may start the journey in the focal platform and end it on another or vice versa. Second, these interactions do not necessarily need to occur within a single session but might comprise several sessions occurring at different times or days. Third, the first-party customer journey comprises the interactions that a customer has with the focal firm aimed at satisfying a *specific need* related to a particular consumption opportunity and/or purchase in a specific category. For example, a customer who interacts with a food delivery platform to get lunch on a Tuesday and dinner on a Friday night is aiming at satisfying two distinct needs on different consumption occasions. Hence, we separate these interactions into two separate journeys.

With this conceptualization of a journey, we develop a probabilistic machine learning modeling framework that allows firms to combine several sources of 1PD across multiple journeys — both present and past journeys — and across many customers. The model flexibly combines multiple types of behaviors, namely search queries, clicks, filters, and purchases, and can accommodate other behaviors collected by the focal company via their website/app or other channels. Using this framework, a focal firm can leverage its 1PD to dynamically infer what a particular customer is looking for as they advance in their focal journey.

2.2 General model specification

We index customers by $i \in \{1, \ldots, I\}$, and their journeys by $j \in \{1, \ldots, J_i\}$, where J_i is the number of journeys customer i has undertaken. We use $t \in \{1, \ldots, T_{ij}\}$ as the cardinal order of actions (hereafter, step) of customer i in journey j. Then, we specify a series of (probabilistic) models, $b = 1, ..., B$ for the behaviors observed along the journey. That is, for customer i in journey j and step t, we define

$$y_{ijt}^{b} \sim g(\cdot|\boldsymbol{\omega}_j, \mathbf{x}_{ijt}^{b}, \boldsymbol{\beta}_{ij}), \tag{1}$$

where g captures the probabilistic relationship between the set of journey-specific query parameters ($\boldsymbol{\omega}_j$), the customer- and journey-specific preferences ($\boldsymbol{\beta}_{ij}$), the observables ($\mathbf{x}_{ijt}^{b}$), and the customer i's behavior b, in step t of journey j.[1] Some behaviors are modeled at each step t (e.g., customers can click at any step), whereas other behaviors are modeled once per journey (e.g., customers insert a query only at the beginning of the journey).[2]

[1] Even though we make distributional assumptions on each component of our model, our framework is general enough to accommodate alternative parametric specifications.

[2] To simplify the notation, we omit global parameters per behavior (e.g., controls and error variances), but these are fully detailed in Web Appendix A.

We model four types of behaviors: queries, clicks, filters, and purchases. Queries are defined as a vector of search query variables $\mathbf{q}_{ij} = \left[q_{ij1}, \ldots, q_{ijM}\right]'$, where each component q_{ijm} corresponds to a different observed variable (e.g., dates of travel and number of passengers in online travel). Because these pieces of information are provided by the customer to obtain a set of product results that match their preferences, we treat each query variable as an outcome that depends on some unobserved component that captures the customer's need in the focal journey. Doing so allows the model to easily account for missing query variables, or query variables that are not present in certain journeys. The query variables could flexibly be binary, categorical, continuous real-valued, or continuous positive-valued. We model these query variables conditionally independently given the vector of journey-specific parameters $\boldsymbol{\omega}_j$ by

$$p(\mathbf{q}_{ij}|\boldsymbol{\omega}_j) = \prod_{m=1}^{M} p(q_{ijm}|\omega_{jm}). \tag{2}$$

Clicks, filters and purchases are modeled jointly in two phases: first, customers explore products through clicks and potential filtering to form a consideration set, and then, customers either choose an item from their considered set or decide not to make a purchase. All of these decisions are guided by a shared set of customer preferences, denoted as $\boldsymbol{\beta}_{ij}$.

We model click choices (y^c_{ijt}) as a discrete multinomial choice made at the page level conditional on customer-journey preferences $\boldsymbol{\beta}_{ij}$ and the set of product attributes $\mathbf{x}^c_{ijtk}$ of each product k shown at step t of journey j,

$$p\left(y^c_{ijt} \mid \boldsymbol{\beta}_{ij},\ \{\mathbf{x}^c_{ijtk}\}_k\right). \tag{3}$$

The choice set is composed of all products shown on the page at step t and two outside options: keep searching to get a new set of products, or finish the search process and move to the purchase decision stage.[3]

We model filter decisions $\mathbf{f}_{ij} = \left[f_{ij1}, \ldots, f_{ijL_{ij}}\right]'$ on a set of L_{ij} attribute levels (e.g., a specific airline like United, or an attribute like non-stop), where each component $f_{ij\ell}$ captures whether the customer filters on attribute level ℓ over the course of the journey.[4] Filters are modeled as binary choices on each level conditionally independent given customer-journey preferences

$$p(\mathbf{f}_{ij}|\boldsymbol{\beta}_{ij}) = \prod_{\ell=1}^{L_{ij}} p(f_{ij\ell}|\boldsymbol{\beta}_{ij}). \tag{4}$$

After clicking and (possibly) filtering through product results, customers make the purchase decision (y^p_{ij}). For the purchase decision, we define the consideration set

[3] We control for ranking effects on search following Ursu (2018) by incorporating the log of the position of product k within the results page in $\mathbf{x}^c_{ijtk}$.

[4] Filters are observed much less frequently than clicks; therefore, we model whether the customer uses a particular filter *at any time* during the journey, rather than at each step t.

$\mathcal{C}_{ij}$ as the set of products that have been clicked on at least once during the course of the journey. We model purchase as a discrete choice among the alternatives in the consideration set $\mathcal{C}_{ij}$ plus the outside option of not purchasing

$$p\left(y_{ij}^{p} \mid \boldsymbol{\beta}_{ij},\ \mathcal{C}_{ij},\ \{\mathbf{x}_{ijk}\}_{k\in\mathcal{C}_{ij}}\right). \tag{5}$$

Clicks, filters and purchase decisions are all driven by customer-journey product-attribute preferences $\boldsymbol{\beta}_{ij}$, whereas queries are solely driven by journey-specific parameters $\boldsymbol{\omega}_{j}$. (Please refer to Web Appendix A for a detailed description of each model component.)

2.3 Heterogeneity: The role of contexts

To facilitate the integration of information within and across journeys, the model accommodates heterogeneity in $\boldsymbol{\omega}_{j}$ and $\boldsymbol{\beta}_{ij}$. In settings involving repeat purchases of frequently purchased items like consumer packaged goods, it is typically assumed that a customer's preferences are stable across journeys (e.g., Rossi et al., 1996, Allenby and Rossi, 1998), and are represented by an unobserved individual-specific component.[5] However, in scenarios such as ours, customers often display journey-specific preferences for journeys involving, for example, domestic versus international flights or traveling for business versus leisure. This variability in preferences complicates the integration of journey data across journeys. To address this issue, our model introduces *journey contexts,* which are latent components that represent specific types of journeys with shared needs across customers.

Methodologically, incorporating context-specific preferences presents several challenges. First, these journey contexts are unobserved and therefore must be inferred from the data. Second, unlike models such as hidden Markov models, there is no systematic transition from one context to another, complicating the identification because the behavior in a previous journey may not predict the current context. Third, the exact number of contexts would likely vary by setting, and therefore we want it to be determined from the data. Lastly, we want to provide enough flexibility to the model such that it will be able to capture *meaningful* contexts as informed by both the query and the customer click-stream behavior (e.g., a "summer family trip" context that bundles together journeys that are more likely to be international trips, with more than one adults and with children, which may involve strong preferences for non-stop destinations and moderate price sensitivity).

To overcome these challenges, we model the journey context as a non-parametric latent segmentation over journeys across customers, using information from the query variables ($\boldsymbol{\omega}_{j}$) as well as the preferences of these journeys that drive clicks, and purchases ($\boldsymbol{\beta}_{ij}$). Specifically, we separate $\boldsymbol{\beta}_{ij}$ into an individual customer heterogeneity

[5] An exception is Jacobs et al. (2016), which posits that different motivations influence varying basket selections.

component and a context heterogeneity component that varies from one journey to another as

$$\boldsymbol{\beta}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\rho}_j, \tag{6}$$

where $\boldsymbol{\mu}_i$ is an individual-specific vector that is shared across all journeys of a particular customer, and $\boldsymbol{\rho}_j$ is a context-specific vector that "shifts" customers' preferences depending on the need of the specific journey but is shared by all customers with the same need.[6]

We account for *customer heterogeneity* by modeling the individual specific vector of parameters following a multivariate normal distribution:

$$\boldsymbol{\mu}_i \sim \mathcal{N}(0, \Sigma). \tag{7}$$

We further define $\boldsymbol{\gamma}_j$ as the vector of all context-specific parameters,

$$\boldsymbol{\gamma}_j = \left[\underbrace{\boldsymbol{\omega}_j}_{\text{Query}} \quad \underbrace{\boldsymbol{\rho}_j}_{\text{Clicks, filters, and purchase}} \right] \tag{8}$$

which includes the set of parameters that determine behaviors in (1).

We model *context heterogeneity* non-parametrically assuming that the context-specific component of a journey, $\boldsymbol{\gamma}_j$, is drawn from an unknown discrete distribution F, which we call the context distribution. We assume that F is drawn using a Pitman-Yor Process. The Pitman-Yor Process is a distribution over infinite almost surely discrete measures used in non-parametric Bayesian models (Pitman and Yor, 1997). Thus, we draw the context-specific parameters, $\boldsymbol{\gamma}_j$, from the context distribution F, and we place a Pitman-Yor process prior on the context distribution F. That is,

$$\boldsymbol{\gamma}_j \sim F, \tag{9}$$

$$F \sim \text{PY}(d, a, F_0), \tag{10}$$

where $0 \leq d < 1$ is a discount parameter, $a > -d$ is a strength parameter, and F_0 is a base distribution over the same space of $\boldsymbol{\gamma}_j$, such that F_0 is the mean distribution of F.

Note that when $d = 0$, the Pitman-Yor process reduces to a Dirichlet process with concentration parameter a and base distribution F_0. The addition of the parameter d allows the drawn distributions from a Pitman-Yor process to exhibit a more flexible long-tail distribution of weights for the mass points, as opposed to the weights decaying exponentially when drawn from Dirichlet processes. This means that the Pitman-Yor process allows for more distinct mass points in the drawn distribution to appear as new observations come in. This feature of the Pitman-Yor process allows the model to capture new contexts that may not have been observed before or contexts that may

[6] Note that because most customers are observed for a few journeys, often a single journey, we cannot directly include a customer-journey-specific term.

happen rather infrequently. In Fig. 1, we show that as more observations come in, the expected unique number of clusters (i.e., contexts in our model) grows for both the Dirichlet process (left most figure in Fig. 1) and the Pitman-Yor process with varying values of the discount parameter d. In contrast to the Dirichlet process, the Pitman-Yor process allows for more flexible patterns of how these unique clusters/contexts appear in the data. Moreover, using a Pitman-Yor process as a prior for our context distribution, similar to the Dirichlet Process, allows our model to infer the number of contexts directly from the data.

We express the context distribution F in terms of the stick-breaking representation of the Pitman-Yor process (Ishwaran & James, 2001), $F = \sum_{c=1}^{\infty} \pi_c \delta_{\theta_c}(\cdot)$ where $\delta_\theta(\cdot)$ is the indicator function and

$$\theta_c \sim F_0, \tag{11}$$

$$\pi_c = V_c \prod_{h=1}^{c-1}(1 - V_h), \qquad V_c \sim \text{Beta}(1 - d, a + c \cdot d). \tag{12}$$

We can then rewrite (9), using the location vectors θ_c and a context assignment variable $z_j \in \{1, 2, \ldots\}$ such that

$$\begin{aligned} \boldsymbol{\gamma}_j &= \theta_{z_j} \\ p(z_j = c) &= \pi_c. \end{aligned} \tag{13}$$

These context assignment variables are useful in understanding the journey-specific need of each journey.

We provide some intuition on how the Pitman-Yor process prior in this model captures the contexts non-parametrically, which we illustrate using Fig. 2. Each F drawn from the Pitman-Yor process is a distribution over contexts, where each location $c = 1, 2, \ldots$ represents a different context (e.g., the summer family vacation, an east-coast business week trip). For any new journey that a customer undertakes, the model draws its journey-specific parameter, $\boldsymbol{\gamma}_j$, from this distribution of contexts, F. This

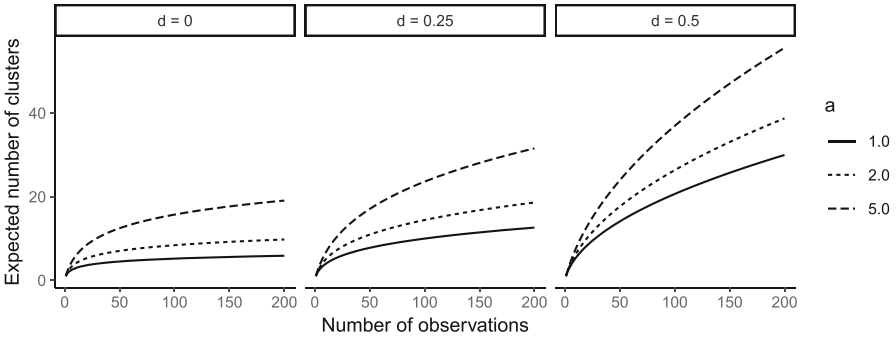


Fig. 1 Expected number of clusters/contexts from a Dirichlet Process ($d = 0$, left) vs. a Pitman-Yor process ($d = 0.25$, middle; and $d = 0.5$, right)

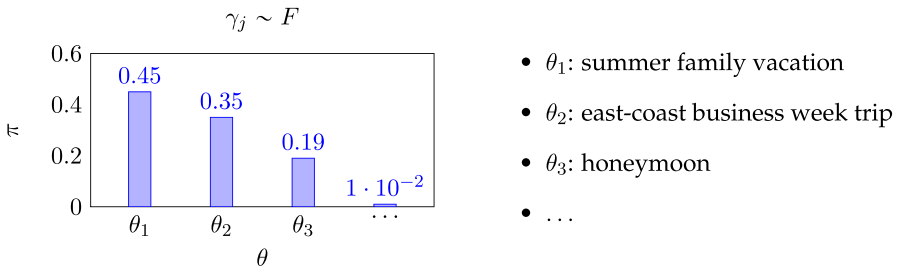


Fig. 2 Example of a context distribution drawn from a Pitman-Yor process prior

drawn distribution is characterized by two main sets of parameters, the locations θ_c and the context sizes π_c. The locations θ_c indicate the set of query, click, and purchase preferences that are associated with context c. The context size π_c represents how likely is context c to be drawn.

We note that due to the infrequent purchase behavior in our data, in contrast with other purchase behavior mixture models (particularly the LDA implementation by Jacobs et al., 2016, where "topic" distributions are individual specific), we assume a common context distribution across customers. In settings with more frequent journeys, one could extend our framework to include customer-specific contexts. That being said, our model captures customer-specific preferences through $\boldsymbol{\mu}_i$.

Next, we describe the sources of information that inform the different parameters of the model. We discuss the identification of the model parameters in Section 3.4, after we describe our empirical setting.

2.4 Sources of information in the model

Our model combines multiple flows of information (queries, clicks, filters, and purchases) from different sources (past and current journeys) to learn what customers might be looking for on different purchase occasions. Our model naturally weights the different pieces of information through the likelihood and the different components in the model, which we describe next. Figure 3 depicts the main assumptions about the data-generating process (for simplicity, the figure omits dispersion across customers, model priors, and filter behavior), which helps to understand how the model fuses information within and across journeys to infer what customers are looking for.

Consider a customer i in a journey j (focal journey) where we have observed t steps that include clicks (and filters), if any. As an ongoing journey, purchase, y_{ij}^p, has not been observed yet. How does the model leverage multiple sources of information to understand the customer's need in this particular journey? First, the model learns from the information provided by customer i during the course of the focal journey j. The queries inform to which context this focal journey most likely belongs to. Queries are a function of query parameters ω_j; which, in turn, are determined by z_j, the context the journey belongs to. The context indirectly informs preferences, $\boldsymbol{\beta}_{ij}$, via ρ_j. Clicks (and filters) provide a strong signal of what the customer wants, as both outcomes

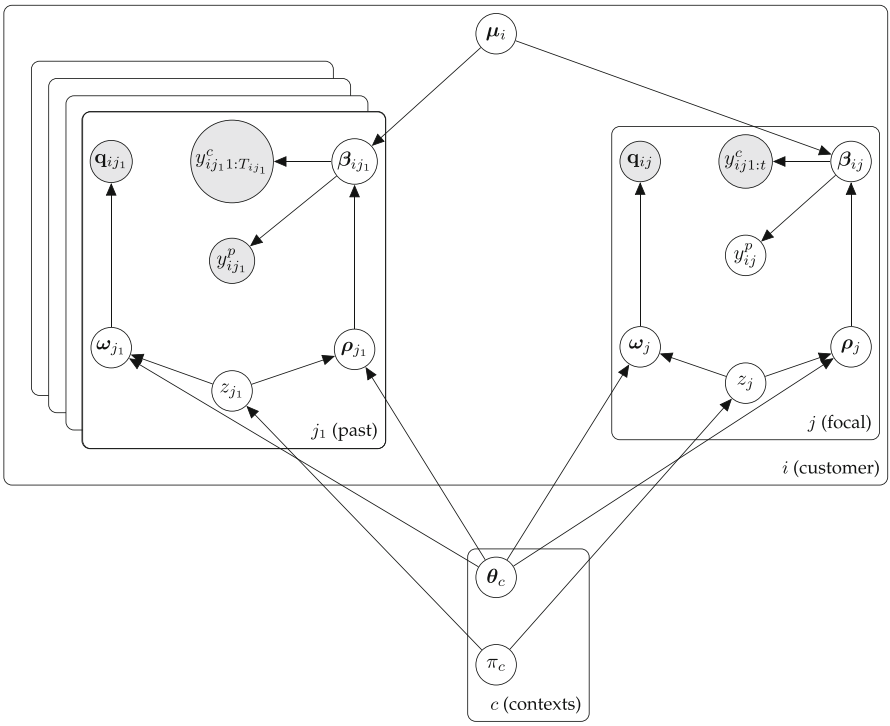


Fig. 3 Simplified directed acyclical graph of the data generating process. Variables in shaded (white) circles are observed (unobserved). We omit from the graph the dispersion across customers (Σ), the consideration set ($\mathcal{C}_{ij}$), parameters a and d from the Pitman-Yor process, and all the priors. To simplify the notation, we also omit filters, $\mathbf{f}_{ij}$, but for all purposes of this figure, we can consider them as part of the clicks, $y^c_{ij1:t}$. Parameters $\boldsymbol{\omega}_j$ and $\boldsymbol{\rho}_j$ are known (deterministic) given the context assignment, z_j, and the context locations $\{\boldsymbol{\theta}_c\}$ (analogous for $\boldsymbol{\omega}_{j_1}$ and $\boldsymbol{\rho}_{j_1}$). Similarly, the vector of preferences, $\boldsymbol{\beta}_{ij}$, is also deterministic given $\boldsymbol{\rho}_j$ and $\boldsymbol{\mu}_i$ ($\boldsymbol{\beta}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\rho}_j$)

are a direct function of preferences, $\boldsymbol{\beta}_{ij}$ (Eqs. 3 and 4). These two assumptions allow our model to integrate information across these different types of actions. Moreover, clicks in the focal journey also inform about the products being currently considered in the journey.

Second, the model learns from queries, clicks, filters and purchases made in past journeys. Those clicks, filters and purchases are all driven by the corresponding preferences of those journeys, all of which share a customer-specific component, $\boldsymbol{\mu}_i$, with the preferences for each particular journey. In addition, past queries are related to the contexts of those journeys because the context shares explanatory power with the customer stable preferences, $\boldsymbol{\mu}_i$, to describe clicks, filters, and purchases in those past journeys. That is, even though past contexts, z_{j_1} and customer stable preferences, $\boldsymbol{\mu}_i$, are unconditionally independent, they are not independent when conditioning on past clicks and purchases, as these are co-determined by past context and stable preferences jointly. Consequently, past queries through past contexts are related to the preferences of the focal journey. Our model "weights" past journeys naturally from the degree to

which these past journeys are related to choices in the focal one. If past journeys are highly informative of the current journey preferences, a larger source of variation of β_{ij} comes from μ_i. Alternatively, if past journeys are less informative, the contexts will drive the larger share of variation in preferences.

Third, the model learns from other customers whose journeys belong to the same context as the focal journey. As contexts are unobserved, the assignment of contexts is probabilistic and the information in other journeys will be shared greatly for those journeys that are more likely to belong to the same context as the focal journey. This information allows the model to learn the context location parameters, $\boldsymbol{\theta}_c$.

2.5 Model estimation and inference: Predicting future activity

The model is estimated on historical data of finished journeys (which may or may not have ended up in a purchase). Once the model has converged, we can easily draw from the posterior to summarize the relevant parameters, including contexts and customer preferences. Importantly, we also want to predict purchase behavior for "ongoing" journeys, defined as those in which the customer might still be exploring/considering to purchase. The challenge in these cases is that consideration sets are not fully observed (as the customer can continue clicking for a while), and thus predicting purchases requires us to augment products that *would be considered* before we can predict the eventual purchase.

Consider an "ongoing" journey j where we observe the query ($\mathbf{q}_{ij}$), clicks ($y^c_{ij1:t}$) and filters up to step t (which we denote as $\mathcal{L}_{ijt} \subseteq \{1, \ldots, L_{ij}\}$). The posterior purchase probability of journey j can be written as

$$p(y^p_{ij}|\mathbf{q}_{ij}, y^c_{ij1:t}, \mathcal{L}_{ijt}) = \int_{\boldsymbol{\beta}_{ij}} p(y^p_{ij}|y^c_{ij1:t}, \boldsymbol{\beta}_{ij}) \cdot p(\boldsymbol{\beta}_{ij}|\mathbf{q}_{ij}, y^c_{ij1:t}, \mathcal{L}_{ijt}) \cdot d\boldsymbol{\beta}_{ij}, \quad (14)$$

where $p(y^p_{ij}|y^c_{ij1:t}, \boldsymbol{\beta}_{ij})$ is a conditional purchase probability given preferences and clicks up to step t, and $p(\boldsymbol{\beta}_{ij}|\mathbf{q}_{ij}, y^c_{ij1:t}, \mathcal{L}_{ijt})$ is the posterior distribution of the preferences given the partial information.

There are two probabilities inside the integral. Computing the latter term is straightforward as it comes directly from drawing from the posterior distributions of our model (see details in Web Appendix D). However, computing the former quantity is more difficult because, to predict a purchase, we need to condition on the customer's consideration set, which might evolve *after* step t. We overcome this challenge by constructing an *augmented* (probabilistic) consideration set that includes products that may be clicked on for the remainder of the journey (i.e., after t). Specifically, we compute

$$p(y^p_{ij}|y^c_{ij1:t}, \boldsymbol{\beta}_{ij}) = \int_{\mathcal{C}_{ij}} p(y^p_{ij}|\boldsymbol{\beta}_{ij}, \mathcal{C}_{ij}) \cdot p(\mathcal{C}_{ij}|y^c_{ij1:t}, \boldsymbol{\beta}_{ij}) \cdot d\mathcal{C}_{ij}, \quad (15)$$

where the first term is computed directly from the model specification, and the second term can be approximated by drawing the consideration set probabilistically. We approximate this second term by a Monte Carlo simulation where we draw each product's membership to the consideration set from a posterior consideration probability $p(k \in \mathcal{C}_{ij} | y^c_{ij1:t}, \boldsymbol{\beta}_{ij})$. If the product k has been clicked before step t, the probability to be considered is one. If not, we approximate it by a reduced-form XGBoost predictor function $\hat{g}_{\mathcal{C}}(\mathbf{x}, \boldsymbol{\beta})$, as this probability is otherwise untractable for a journey that has not finished, and thus, we do not observe how many more steps the journey will have. We train such predictor function using journeys from the training data and $\boldsymbol{\beta}$ drawn from the posterior distribution. We describe this approach in detail and outline the full procedure to compute purchase probabilities in Web Appendix E.

2.6 Alternative modeling approaches

Our primary objective is to dynamically infer customer preferences at the journey level as the journey unfolds, thereby enhancing personalized marketing strategies. For these strategies to be effective, managers require a methodology with sufficient predictive validity to accurately infer consumer preferences and act on them. We discuss alternative possible approaches to achieve this goal.

One could consider the extensive literature on search models, particularly those based on the sequential (Weitzman, 1979) search model (e.g., Kim et al., 2010, Koulayev, 2014, Honka and Chintagunta, 2017), which offer valuable insights and counterfactual evaluations. Although accommodating heterogeneous preferences at the population level is relatively straightforward, integrating such heterogeneity into a structural search model to obtain individual-level estimates and updating them as new information arrives has historically posed challenges. These models are computationally intensive to estimate and struggle to scale with large datasets (a notable exception being Morozov, 2023). This challenge is amplified when modeling multiple journeys across customers, as the requisite heterogeneity specification must pool information across journeys within and between customers, significantly increasing computational demands.[7] Inferring preferences at the journey level is essential for tailoring marketing strategies like retargeting campaigns or product promotions during ongoing searches. By relaxing structural assumptions (particularly those derived from click ordering), our model offers both higher flexibility to better match the observed patterns in the data and computational advantages in both estimation and inference for ongoing journeys.

Moreover, our model seamlessly integrates various types of information: search queries, filters, clicks, and purchases. While filters have been incorporated into structural search models (e.g., Chen and Yao, 2017), to the best of our knowledge, our approach is the first to include search queries in modeling clicks and purchases. This probabilistic modeling technique enables us to better infer granular preferences at the journey level by drawing insights from similar journeys of other customers. One limitation of our modeling approach relative to the structural search modeling approach

[7] This is the main reason why we do not explicitly compare our model's predictive performance against a structural search model.

is that our approach cannot derive policy invariant parameters for counterfactual scenarios that significantly alter market equilibrium, such as airline mergers impacting flight offerings or significant changes in how information is presented to consumers affecting search costs. For counterfactual analysis in such scenarios, structural search models may be more suitable.

Another alternative route would be a fully predictive approach using ML models such as a neural network. While these models might meet predictive validity goals, they often lack in providing a deep understanding of customer preferences over attributes, limiting their applicability to tasks like cross-selling campaign responses. Marketers often need to decode the black box of ML models to recommend products in different categories than those the model was trained for. Furthermore, traditional ML models often do not capture unobserved customer differences as effectively as hierarchical models and require significant feature engineering, such as selecting which moments to summarize click and purchase histories. For instance, Smith et al. (2023) demonstrate that hierarchical choice models outperform random forest and XGBoost models in leveraging purchase history for targeted pricing. Conversely, our approach models a comprehensive set of behaviors within a single framework, naturally incorporating past journeys through posterior distributions. This enables continuous prediction throughout the customer journey without needing to fully retrain the model. In Section 4.4, we compare the predictive performances of our model for ongoing journeys relative to several ML approaches (see further details in Web Appendix H). Our findings show that across prediction tasks at different stages of the customer journey, our model performs more robustly than the best ML alternatives.

To summarize, our model strikes a balance between flexibly allowing for within and across journey heterogeneity and model interpretability. This allows us to facilitate managerial decisions, such as retargeting, cross-selling, and product recommendation, where the space of the products is very large and dynamic. Additionally, understanding customer needs beyond predictions aids in optimizing website design to better meet those needs (see Section 5.2).

3 Empirical setting

We apply our model to the context of airline ticket purchases using data from one of the largest worldwide online travel platforms. The dataset contains click-stream data on the focal platform of 4,500 customers who searched for flight tickets between May 2017 and November 2017.[8] For each web page shown to those customers, we observe the customer ID, the timestamp of when the customer accessed the page, the search query inputs associated with that page, and the list of results (including the flight attributes such as price, length, or airline carrier) observed by the customer after entering the query. The data also contains clicks on specific flights and the confirmation page for journeys that ended up in a purchase. We observe a total of 5,285,770 flight

[8] The focal firm randomly selected 4,500 customers among those that they define as "active" users.

offers displayed in 106,018 results pages, which resulted in 3,718 flight itineraries purchased.

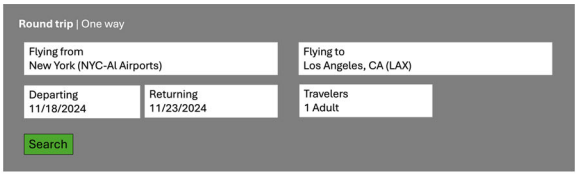
3.1 The first-party journey of airline travel

Consistent with our conceptualization of a first-party journey, the journey starts when the customer lands at the website's homepage to search for a flight. There are two types of trips that the customer can choose from: (1) Roundtrip, and (2) One-way.[9] For roundtrips, the customer includes an origin and a destination; a departure date for the portion of the trip from the origin to the destination, known as the *outbound leg*; and a returning date for the portion of the trip from the destination back to the origin, known as the *inbound leg*. Each leg of the trip is composed by either one non-stop flight or multiple connecting flights. One-way itineraries have only one direction of travel.

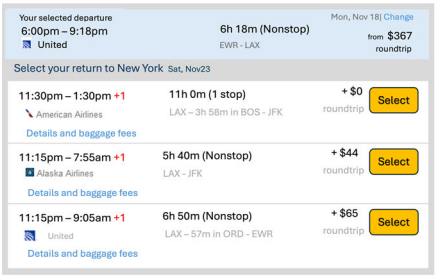
Note that, as is the case in numerous business settings, a customer journey can be highly non-linear (Grewal & Roggeveen, 2020). That is, the customer may go back from each step to enter a new/revised query, to click on alternative outbound or inbound results, etc. Moreover, this process does not need to occur during the same internet session, but can occur over the course of multiple days (Lee et al., 2018). Accordingly, we use the queries in our data to construct a flexible definition of the customer journey by combining pages/sessions that belong to the same customer with the same "trip need." Specifically, a customer journey comprises all sessions that, while they might have occurred at different points in time, (1) have departing or arrival dates within up to 4 days; and (2) have origin or destination to close-by airports and cities within a 140 miles range (approx. 225 km.). Once all these pages are combined, we sort them by timestamp and remove subsequent searches of the same journey after a purchase is made, to remove the infrequent behavior of customers checking prices of the same itinerary after purchase. To avoid mislabeling censored and potentially unfinished journeys as no-purchase journeys, we remove all journeys (where the purchase has not been observed) in which the itinerary's first flight departs after our observation window's end. This process resulted in a total of 25,402 journeys, corresponding to an average of 5.6 journeys per customer. The conceptualization of journeys as described above, rather than simply using individual search queries as sessions, allows us to seamlessly integrate behavior across sessions that are aimed at covering the same need. Next, we describe the flow of the roundtrip purchase journey, since the one-way purchase journey is a nested version of the roundtrip purchase journey.

After a customer lands on the homepage, they start specifying the search query (see Fig. 4a) by selecting the type of trip to search for (e.g., *roundtrip*) and filling multiple fields (all of them required): origin and destination cities/airports, outbound and inbound departing dates (i.e., "departing" and "returning dates" in Fig. 4a, respectively), and travelers (number of adults and children). The customer then clicks on the "Search" button, which triggers the platform to search the flight results that match the

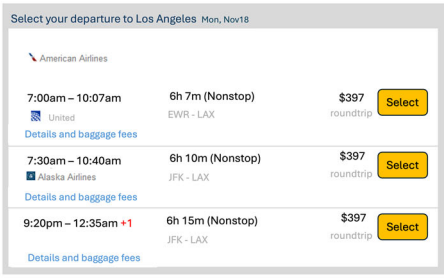
[9] We drop from our analysis the third type of trip, multi-cities trips, as they constitute a very small portion of the trips. Moreover, our data does not contain searches on packages (e.g., flight + hotel), and therefore we focus on journeys over flights only.



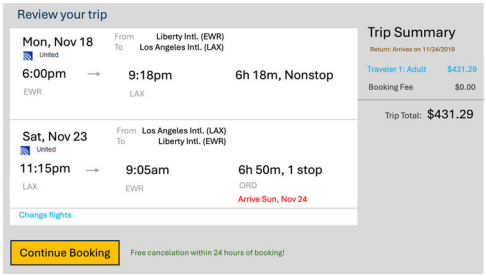
(a) Example of a query page



(b) Example of an outbound page results



(c) Example of an inbound page results



(b) Example of a flight details results page

Fig. 4 Purchase journey steps

information from the query. The website displays the set of results, sorted increasingly by price, for the outbound itineraries (see Fig. 4b). Each of these itineraries are fully described by a path of flights that start at the origin airport and finish at the destination airport. The website clearly displays all relevant information of the outbound legs of the search results, including price, the total duration of leg, the airline carrier, the number of stops, and departure/arrival times. Note that for the outbound leg, the price displayed corresponds to the price of the complete roundtrip itinerary, including the price of the outbound leg and the cheapest inbound leg that corresponds to the outbound leg. At this point, the customer might want to explore some of the presented options (see "Select" below), click "More" to get exposed to more flights, "Filter" to restrict the characteristics of the flights to be shown, or abandon the search.

If the customer clicks on the "Select" button of one of the outbound offers, the website displays the set of corresponding inbound results that correspond to the clicked outbound leg (see Fig. 4c). For those resulting inbound offers, the website displays the same level of information displayed for the outbound offers (see Fig. 4c), including

the extra price of each alternative compared to the minimum price (i.e., the price displayed in the outbound page of results). Once the customer clicks on the "Select" button of one of the inbound offers, the website shows a page with the details of all the information mentioned before from both the outbound and the inbound legs (see Fig. 4d), as well as the full breakdown of the price (taxes and fees clearly displayed). After the customer clicks on "Continue Booking", the customer fills in information about the passengers and proceeds with the payment steps. Finally, after finalizing the purchase, the customer is shown a confirmation page. The one-way purchase journey is very similar, with the exception that instead of clicking through two sets of results (outbound and inbound), the customer is displayed only one page of results, "One-way results".

3.2 Empirical model specification

In this subsection, we discuss how to adapt our model from Section 2 to the travel platform application.

3.2.1 Search queries

We construct several variables from the query information that aim to capture the context of a journey in our model. While some pieces of information are directly provided by the customer (e.g., destination), others can be indirectly determined; e.g., whether the trip includes weekends which can be extracted from the dates, or the trip distance (inferred from the origin and destination airports). We combine these variables into a vector of query variables, $\mathbf{q}_{ij}$ in (2), that aim to capture information about the journey in four different dimensions: (1) who is traveling, (2) which market this flight belongs to (origin-destination), (3) when is the trip, (4) when was the search made. We model each of these variables with different distributions depending on their type: Bernoulli for binary variables, Categorical for categorical ones, and exponential and Gaussian for continuous variables. Table 1 shows these variables and their corresponding summary statistics.

We observe a great variety of trip characteristics in the data: 66% of journeys are roundtrip (vs. one-ways); 28% include more than one adult and 8% include kids. The average stay for roundtrips is 11.80 days, 37% of journeys are for flights during the summer season and 3% for the holiday season,[10] and 66% of flight searches include stays over weekends. The average trip distance is 3,548 kilometers or 2,205 miles (e.g., approx. New York to Las Vegas); 59% of journeys are domestic (including US-Canada, within-EU, or within-country flights) and 51% are US Only. Purchase journeys occur, on average, 50.73 days prior to the departing date; 92% introduce a departing location code for an airport (e.g., JFK), 8% a departing code of a city (e.g., NYC), and the rest include a departing code that refers to both city and airport (e.g., MIA).

[10] We define the summer season from June 30th to September 4th, and holiday season stays that include either Thanksgiving, Christmas, or New Year's holidays.

Table 1 Summary statistics of query variables

Query variable	Mean	SD	Quantiles		
			5%	50%	95%
Continuous					
Trip distance (km.)	3,584.16	3,465.07	448	2,269	11,529
Time in advance to buy (days)	50.73	59.82	1	29	182
Length of stay (only RT) (days)	11.80	21.25	2	6	37
Binary					
Is it roundtrip?	0.66	.	0	1	1
Traveling with kids?	0.08	.	0	0	1
More than one adult?	0.28	.	0	0	1
Is it domestic?[a]	0.59	.	0	1	1
Is it summer season?	0.37	.	0	0	1
Holiday season?	0.03	.	0	0	0
Does stay include a weekend?	0.66	.	0	1	1
Flying from international airport?	0.74	.	0	1	1
Searching on weekend?	0.21	.	0	0	1
Searching during work hours?	0.49	.	0	0	1
Categorical					
Market					
US Only	0.51	.	0	1	1
US Overseas	0.18	.	0	0	1
Within North America[b]	0.15	.	0	0	1
Non-US within continent	0.10	.	0	0	1
Non-US across continent	0.06	.	0	0	1
Type of departure location					
Airport	0.92	.	0	1	1
Multi-airport City	0.08	.	0	0	1

[a] We define domestic as flights between the US and Canada, as well as flights within the European Union (EU)
[b] This category includes Canada and Mexico and excludes US-only trips

3.2.2 Clicks and purchase

Once the query information is captured, we build a set of "click occasions" faced by the customer. These click occasions are composed of a set of alternatives to click on and the outcome of what was actually clicked on (or not). There are two types of click occasions: (1) those on an outbound results page (where clicking on a product leads to an inbound results page) and (2) those on an inbound results page (where clicking on a product leads to flight detail page). We define Page_{ijt} as the set of products displayed to customer i in journey j at step t. We observe and allow in our model customers to click on multiple flights from each results page by adding a click opportunity with the same page results for each additional clicked product. By default, results within a page

are sorted increasingly in price. Once the customer sees the results, they can sort the flights differently by clicking on sorting options, which include leg duration (hours) and departure/arrival time. While these alternative sorting actions could be valuable pieces of information for inferring preferences, unfortunately, we do not observe them explicitly because the firm records these actions as if the customer starts the search again, which is how we model them.

We specify the click component of our model in (3) as a discrete choice model, where customers can choose between clicking on a product shown on the page $k \in \text{Page}_{ijt}$, continue searching to get a new set of products ($k = s$), or finish the search process and move to the purchase decision ($k = e$), which could mean either purchasing a considered product or deciding not to buy. We use a multinomial probit specification with latent propensities u^c_{ijtk}, such that

$$y^c_{ijt} = \underset{k \in \text{Page}_{ijt} \cup \{s,e\}}{\arg\max} \left\{ u^c_{ijtk} \right\}, \text{ with}$$

$$u^c_{ijtk} = \begin{cases} \beta^{0c}_{ij} + \mathbf{x}^{c}_{ijtk}{}' \cdot \boldsymbol{\beta}^x_{ij} + \text{log-rank}_{ijtk} \cdot \eta + \varepsilon_{ijtk} & \text{if } k \in \text{Page}_{ijt}, \\ \beta^{0s}_{ij} + \varepsilon_{ijts} & \text{if } k = s, \\ \beta^{0e}_{ij} + \varepsilon_{ijte} & \text{if } k = e, \end{cases} \tag{16}$$

where $\varepsilon_{ijtk} - \varepsilon_{ijte} \sim \mathcal{N}(0, \sigma^2)$, $\mathbf{x}^c_{ijtk}$ is the vector of attributes of product k, $\boldsymbol{\beta}^x_{ij}$ is the vector of customer- and journey-specific product-attribute preferences, β^{0c}_{ij} is the intercept for clicking on a product, β^s_{ij} is the intercept for the decision to continue searching, and β^{0e}_{ij} is the intercept for finishing the search process, normalized to 0 for identification purposes.

We control for ranking effects on search (Ursu, 2018) by incorporating the log of the position of product k within the results page into the search in u^c_{ijtk} and using η to capture such ranking effects.[11] Such a term also captures search costs within a page, along with the intercepts in (3) that capture users' propensity to keep searching and are related to search costs across pages.

We create the consideration set, $\mathcal{C}_{ij}$ in (5), by including all products that were clicked before purchase (e.g., Bronnenberg et al., 2016)

$$\mathcal{C}_{ij} = \left\{ k \ : \ k \in \text{Page}_{ijt}, \ y^c_{ijt} = k, \ t \in \{1, \ldots, T_{ij}\} \right\}.$$

For roundtrip flights, only products where both the outbound and inbound legs of the itinerary were clicked on are added to the consideration set. We then register the outcome of the purchase occasion as a purchase for the product that was purchased (a purchase confirmation page was presented), if any, or as a non-purchase in case no product was purchased.

We specify the purchase model in (5) as a discrete choice between purchasing a previously considered product and not purchasing. Specifically, we use a multinomial

[11] Following Ursu (2018), we include the log-position rank in the click decision but not in the purchase-given-clicks decision.

probit specification with latent propensities u^p_{ijk}, such that

$$y^p_{ij} = \underset{k \in \mathcal{C}_{ij} \cup \{\text{NoPurchase}\}}{\arg\max} \left\{ u^p_{ijk} \right\}, \text{ where}$$
$$u^p_{ijk} = \begin{cases} \beta^{0p}_{ij} + \mathbf{x}_{ijk}{}' \cdot \boldsymbol{\beta}^x_{ij} + \epsilon_{ijk} & \text{if } k \in \mathcal{C}_{ij} \\ \beta^{0o}_{ij} + \epsilon_{ijo} & \text{if } k = \text{NoPurchase}, \end{cases} \tag{17}$$

with $\epsilon_{ijk} - \epsilon_{ijo} \sim \mathcal{N}(0, \sigma^2_p)$, and where $\mathbf{x}_{ijk}$ is the vector of attributes of product k, $\boldsymbol{\beta}^x_{ij}$ is the same vector of customer- and journey-specific product-attribute preferences as in (16), β^{0p}_{ij} is the intercept for purchasing a product, and β^{0o}_{ij} is the intercept for not buying, normalized to 0 for identification purposes.

Note that the vector of customer- and journey-specific preferences $\boldsymbol{\beta}_{ij}$ introduced in our general model in (1) contains the product attribute preferences as well as the intercepts for the click and purchase models,

$$\boldsymbol{\beta}_{ij} = \left(\beta^{0c}_{ij},\ \beta^{0s}_{ij},\ \beta^{0p}_{ij},\ \boldsymbol{\beta}^{x\,\prime}_{ij} \right)'.$$

3.2.3 Product attributes

Customers observe multiple product attributes when making a click and purchase decision. For a roundtrip journey, all attributes, except price, are specific to *each leg* of the trip. That is, there is a set of attributes that describe the outbound leg of the trip, and there is the same set of attributes that describe the inbound (returning) leg of the trip. Our model allows users to look for different attributes in a flight, depending on the leg.[12]

A subset of these attributes is summarized in Table 2 (see Web Appendix F for full set of summary statistics). Prices are measured at the whole trip level. The average offer displayed is priced at $1,547 but offers vary significantly in their price, with a standard deviation of $3,249. Furthermore, journeys have different price levels that depend on origin-destination and the dates. This variation in price becomes clearer when looking at the price of the cheapest offer per journey. The cheapest price displayed per journey has an average of $698 across all customer journeys, with a standard deviation of $1,526. This indicates that raw prices may not be a good proxy to capture price sensitivity as prices are only compared within a journey. For example, a New York - Chicago roundtrip ticket for $600 may be considered expensive, whereas a roundtrip flight from New York to Buenos Aires for $800 may be considered a good deal. Therefore, we transform prices by computing the log difference between the focal flight and the lowest price per journey.[13] Log transformation accounts for the long-tail dispersion in price differences.

[12] For roundtrips, we assume the same preferences for outbound and inbound attributes, except for preferences for departure and arrival times.

[13] We believe this transformation is appropriate due to the salience of price comparisons in this context. We tested both specifications (raw prices and relative prices) and found that the 'relative log price' provided slightly better predictive performance.

Table 2 Summary statistics of a subset of product attributes in page results

Product attribute	Mean	SD	Quantiles		
			5%	50%	95%
Product level attributes					
Price	1,547	3,269	196	751	5,320
Cheapest price per journey	698	1,526	98	401	2,117
Outbound level attributes					
Length of trip (hours)	11.28	8.49	2.05	8.42	28.60
Shortest length of trip per journey (hours)	5.86	5.05	1.25	4.07	17.08
Number of stops: Non stop	0.20	.	0	0	1
Alliance: OneWorld (American)	0.27	.	0	0	1
Alliance: Skyteam (Delta)	0.27	.	0	0	1
Alliance: Star Alliance (United)	0.23	.	0	0	1
Dep. time: Early morning (0:00am - 4:59am)	0.04	.	0	0	0
Inbound level attributes					
Length of trip (hours)	11.08	9.02	1.83	7.92	29.50
Shortest length of trip per journey (hours)	6.17	5.31	1.25	4.27	17.75
Number of stops: Non stop	0.19	.	0	0	1
Alliance: OneWorld (American)	0.51	.	0	1	1
Alliance: Skyteam (Delta)	0.13	.	0	0	1
Alliance: Star Alliance (United)	0.15	.	0	0	1
Dep. time: Early morning (0:00am - 4:59am)	0.03	.	0	0	0

Offers also differ in terms of how long each leg of the trip is. The average outbound leg of a displayed trip takes 11.28 hours, with a large variation within and across journeys. The shortest flight per journey takes, on average, 5.86 hours for the outbound leg. Most displayed flights are one-stop flights (59%-70% for inbound or outbound legs), whereas nonstop flights account for 19%-20% of offers. As most of the variation in length is driven by the number of stops a leg has, we subtract from the length of a leg the length of the shortest itinerary with the same number of stops to isolate the effect of longer connections from more connections.

Airline data is sparse, so we aggregate airlines into alliances. Alliances are groups of airlines that share benefits and usually operate code-shared flights. For example, a JFK to Madrid flight operated by Iberia might also be sold by American Airlines, British Airways, and Finnair, all belonging to the same alliance. The three biggest alliances are Oneworld, SkyTeam, and Star Alliance, accounting for 77%-79% of all offers. Individual airlines not in an alliance but representing a significant proportion of offers are categorized as their own airline, while smaller airlines not in an alliance are grouped under "Other-No alliance". Finally, we label as "Multiple alliances" offers that have connecting flights of different alliances in the same leg of the trip.

3.2.4 Filters

Customers can filter the product results on a page by clicking on one (or multiple) product attribute levels: airline, number of stops, and departure and arrival times. In the data, we do not directly observe the use of filters. Instead, we infer the application of filters from the data by contrasting the product results shown to the customer at each step of the journey with the set of products available at the beginning of the journey. Number of stops is the attribute with the largest number of filter occasions in our data, with 13.8% of journeys having the customer filtering for non-stop flights at some point during the course of the journey. For all other attributes and levels, filters are applied very infrequently (less than 4% of journeys), supporting the model simplification of modeling filter outcomes once for the entire course of a journey as opposed to at the page-level. Web Appendix F.2 discusses the construction process of the filters and summary statistics of the filtering actions.

We specify each model component $p(f_{ij\ell} \mid \boldsymbol{\beta}_{ij})$ in (4) using a binary probit specification such that

$$f_{ij\ell} \sim \text{Bernoulli}\left(\Phi\left(\alpha_\ell^0 + \mathbf{w}_{ij\ell}{}' \cdot \boldsymbol{\alpha}_\ell^w + \boldsymbol{\beta}_{ij}^{x\,\prime} \cdot \boldsymbol{\alpha}_\ell^\beta\right)\right), \tag{18}$$

where α_ℓ^0 is the intercept of filtering on level ℓ, $\boldsymbol{\beta}_{ij}^x$ is the same set of preferences that drive clicks and purchases, and $\boldsymbol{\alpha}_\ell^\beta$ is the vector that relates those preferences to the filtering decision. It is this term that allows the model to learn preferences for attributes by fusing filtering decisions about those attributes. To control for other factors that may affect the filtering decisions (e.g., the overall characteristics of unfiltered results), we include a set of controls $\mathbf{w}_{ij\ell}$ that summarize the set of (unfiltered) results (the number of total products, the percentage of products with level ℓ, and the number of top 5 products with level ℓ in the unfiltered results).

3.3 Summary statistics

Table 3 shows the the total number of customers, journeys, purchases, click steps and clicked products. We observe a total of 25,402 journeys, for which we aim to estimate individual-level preferences. The data indeed exhibit thin past purchase history at the individual level—while, on average, each customer undertakes 5.645 journeys, the average number of purchases per customer is only 0.83. With low historical purchase rates at the customer level, the use of traditional models that rely on long individual purchase histories, such as scanner panel data, is limited. In many settings such as ours, the 1PD on purchases is thin. However, because customers often use the platform for search — even when a purchase eventually does not occur — these searches can still provide valuable information. The amount of data collected during the journey is rich—on average, customers in our sample clicked on 1.14 products per journey, having a total of 6.45 clicks in total.

On average, 14.6% of journeys end with a purchase. This number may seem high when compared to standard metrics of conversion for an online retailer, but note there are two possible reasons for this relatively high conversion rate. First, our data

Table 3 Data summaries, per customer and per journey

Variable	Total	Average per...							
		Customer		Journey		Purchased journey		Non-purchased journey	
		Mean	s.e.	Mean	s.e.	Mean	s.e.	Mean	s.e.
Customers	4,500	.	.	.	.	.	.	.	.
Journeys	25,402	5.645	0.076	.	.	.	.	.	.
One-way	8,692	1.932	0.047	.	.	.	.	.	.
Roundtrip	16,710	3.713	0.056	.	.	.	.	.	.
Purchases	3,718	0.826	0.016	0.146	0.002	1.000	.	0.000	.
Steps	106,018	23.560	0.352	4.174	0.037	6.909	0.119	3.705	0.037
... in OW search	40,184	8.930	0.262	4.623	0.068	6.127	0.155	4.174	0.074
... in RT outbound	51,393	11.421	0.208	3.076	0.036	5.270	0.141	2.824	0.036
... in RT inbound	14,441	3.209	0.046	0.864	0.012	2.547	0.053	0.671	0.011
Clicked products	29,037	6.453	0.072	1.143	0.014	2.945	0.048	0.834	0.013
... in OW search	5,884	1.308	0.035	0.677	0.013	1.676	0.031	0.379	0.011
... in RT outbound	14,441	3.209	0.046	0.864	0.012	2.547	0.053	0.671	0.011
... in RT inbound	8,712	1.936	0.029	0.521	0.008	1.872	0.036	0.366	0.007
Filtered attributes	10,121	2.249	0.056	0.398	0.006	0.639	0.019	0.357	0.006
... in OW search	4,654	1.034	0.039	0.183	0.004	0.381	0.016	0.149	0.004
... in RT	5,467	1.215	0.035	0.215	0.005	0.258	0.013	0.208	0.005

correspond to a sample of customers defined as "active" by the focal firm, and therefore this figure would be lower for the average website visitor. Second, in this paper, we adopt a broader definition of a journey, the first-party customer journey, which not only includes multiple sessions for the same customer but also combines searches that include nearby airports on similar dates, as the customer is trying to satisfy the same need. In contrast, traditional conversion rates tend to treat different search queries, with different variations of airports or dates, as different and independent purchase funnels.

3.4 Identification and estimation

We now summarize the main data patterns that enable the identification of our model's parameters. First, variations in attributes across alternatives, along with variations in clickthrough and purchase rates, allow for the identification of preferences over product attributes. Second, because results are deterministically sorted by price, and the order is observed, we can disentangle the price coefficient from ranking effects due to the variation of prices across options in the same position and products with the same price in different positions. Third, the different sources of heterogeneity are identified through systematic patterns of preferences. Customer heterogeneity is identified by variations across journeys belonging to the same customer. Similarly, context heterogeneity is identified through systematic multivariate similarities across both query and preference parameters. Finally, the number of contexts is identified through a combination of the degree of similarity between journeys and the sparsity enforced by the Pitman-Yor process priors.

We split the data into training and validation at the customer and journey level: for each customer, we use some of their journeys for training and leave the last journey (or last few journeys) as a hold-out, such that we can explore the model performance in new journeys for existing customers. We leverage the data split in different ways. The training data is used to estimate the model and to summarize overall preferences and contexts in this market (Sections 4.1 and 4.2). We use held-out journeys to illustrate model inferences at different stages of the journey (Section 4.3), to evaluate the model's overall predictive ability (Section 4.4), and finally, to quantify the value of leveraging first-party journeys under different assumptions regarding the traceability of customer data (Section 5).

We estimate the model parameters in a fully Bayesian framework using MCMC. Specifically, we use a blocked Gibbs sampler implemented in Julia to draw from the context posterior distribution with Pitman-Yor process priors following its stick-breaking representation (Ishwaran & James, 2001). This approach allows us to implement a fast sampler that is able to draw the context assignment, z_{ij}, in parallel across journeys, given context locations and context size parameters, as opposed to using marginal samplers that marginalize the context distribution F but sample sequentially for each journey conditioning on context assignments for all other journeys (Neal, 2000). We use adaptive Metropolis-within-Gibbs steps to draw the Pitman-Yor parameters a and d, and we use Gibbs steps to draw the rest of the parameters using their full conditionals. For full details of the estimation procedure see Web Appendix C. We

estimate our model for 100,000 warm-up iterations and we use a sample of 1,000 draws from the posterior distribution (5,000 iterations saving a draw every 5 iterations). We assess convergence by monitoring traceplots over the model parameters.

4 Results

4.1 Average preferences for flight attributes

We start by describing the mean level preferences ($\boldsymbol{\beta}_{ij}$). As these preferences vary across journeys, we compute the population mean estimates averaging $\boldsymbol{\beta}_{ij}$ across all in-sample journeys (Fig. 5). As expected, customers prefer lower prices and flights of shorter lengths, they significantly prefer non-stops over one-stop, and tend to prefer the OneWorld alliance over all other alternatives.[14] They favor departure times either in the morning or in the afternoon for the outbound leg, while they prefer to depart in the afternoon or in the evening for the return leg.

4.2 Contexts uncovered from the data

While the mean parameters are informative to explore the overall preferences for airline tickets, customers book flights to satisfy a wide variety of needs (i.e., a family vacation is not the same as a 2-day conference trip), for which preferences might also vary. We explore those differences by examining the rich set of contexts uncovered by our model. As explained in Section 2 and illustrated in Fig. 1, we recover the number of contexts non-parametrically. Figure 6 shows the posterior distribution of the Pitman Yor parameters as well as the number of contexts (i.e., those with at least one journey assigned to them across the posterior distribution). While Fig. 6 shows a maximum of 39 contexts in the data, when looking at the relative size of each context (Fig. 7), we find that only 22 of these contexts appear in a journey with a probability higher than 1% (dotted line). We focus on those 22 contexts next.

The model parameters allow us to explore the type of trip (or "need") that each context represents. To interpret the model parameters we need to normalize the location parameters, θ_c, to compare both the query parameters and the flight preferences across the 22 contexts (see posterior statistics for all parameters of the 22 contexts in Web Appendix G.1). We normalize the location parameters to compare contexts with respect to whether they score higher or lower than average on each of the query parameters and preferences. Figure 8 shows these relative scores for a subset of variables.[15]

Finally, using airport information, we identify the top 50 most frequent routes per context (see Fig. 9 for a sample of contexts and Fig. G.3 in the Web Appendix for all contexts). Note that the exact destination is not included in the model and therefore not

[14] For categorical attributes, the base levels are: one-stop, OneWorld, and morning times.

[15] For simplicity, we present the results for a subset of query parameters and product-attribute preferences. Details on the normalizing procedure and the full set of results are shown in Web Appendix G.2.

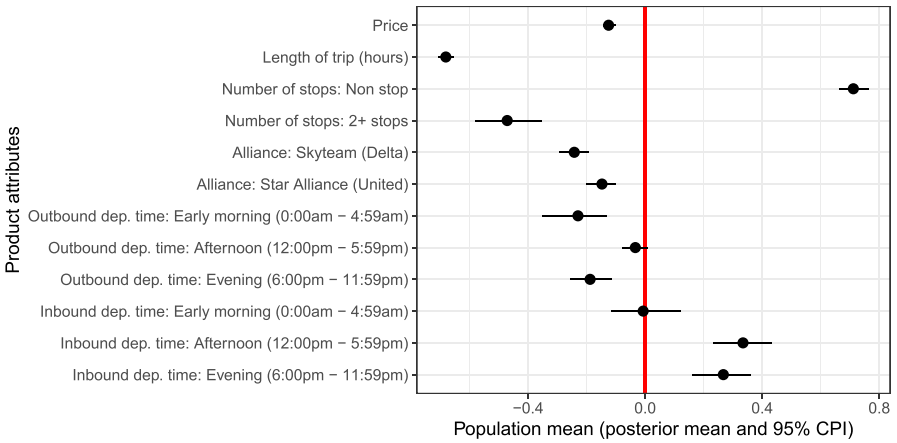


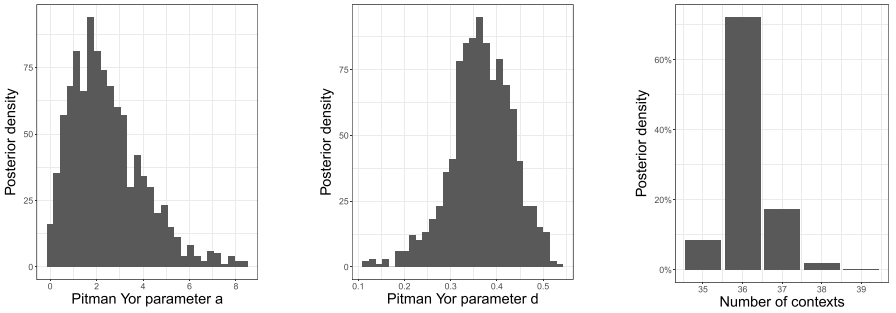
Fig. 5 Population level estimates: Posterior Mean and 95% Credible Posterior Interval (CPI)

used to draw contexts from the data. However, we summarize this information for two main reasons. First, to explore the external validity of the contexts uncovered by the model. We find that destinations are largely congruent with the queries and preferences of each context (e.g., Hawaii is a common destination for week-long family trips with a strong preference for non-stop flights). Second, flight destinations might be useful to further characterize the contexts uncovered by the model.

For example, combining the insights from the model parameters and top destinations, we label Context 1 as "One-way solo domestic trips (mostly US)," as these journeys tend to be one-way and domestic, primarily involving no other adults or children. Searches for these journeys typically begin approximately 24 days before departure. Compared to the general population, customers in these journeys are slightly more price-sensitive, have a stronger dislike for longer flights and routes with two or more stops, and show weaker preferences for routes with Spirit, Frontier, and multiple alliance flights. Similarly, we label Context 5 as "Family vacations in the Caribbean," since a prototypical journey in this context is a roundtrip route between the US and other North American or Caribbean countries. These journeys often involve additional adults and children, with an average stay of 9.8 days and searches beginning around 82 days before departure. There is a pronounced preference for Star Alliance, non-stop flights, and shorter routes.

4.3 Learning *along* the journey

Another desirable characteristic of the model is that it enables the focal firm to update customer insights as users advance in their journey. To illustrate this process, we select a customer with two journeys in the holdout data and examine the model inferences at different stages of each journey. Figure 10 shows the model posterior distribution of contexts (top row) and price sensitivity (bottom row) for each journey, given the data available at four different stages: (1) at the homepage, (2) after the query was inserted, (3) after two click steps, and (4) after five click steps.



(a) Posterior density of a (b) Posterior density of d (b) Posterior density of the number of contexts

Fig. 6 Posterior density of Pitman Yor parameters and resulting number of contexts

At the homepage, the model does not have any information about the focal journey. Hence, the inference for context corresponds to the average propensities across the population. The small differences between the first and second journeys are simply due to sampling error. Then, the model incorporates the query information and updates its inference about the context (first row, second column). For *Journey 1* ("One adult roundtrip from Kyiv, Ukraine to New York, USA"), the probability for contexts 7 and 10 notably increases, whereas for *Journey 2* ("One adult oneway, Kyiv, Ukraine to Lisbon, Portugal") contexts 9, 13, and 20 become more likely to occur. After 2 clicks (steps), the likelihood that *Journey 1* belongs to context 10 increases further, whereas the probabilities for *Journey 2* remain largely unchanged. Finally, after 5 clicks (steps), the model has more information about what the customer is clicking on. Notably, *Journey 1* is clearly identified as belonging to context 10 (which we describe as the 2-week long-haul summer trip), whereas the inference for *Journey 2* changes drastically, with context 1 having the highest probability.

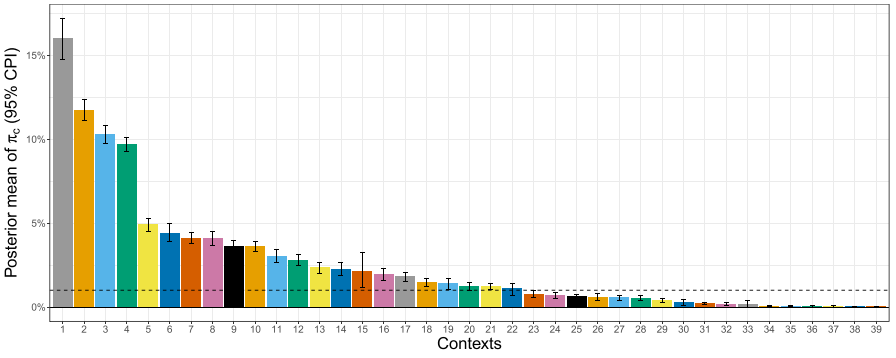


Fig. 7 Posterior mean (and 95% CPI) of context probabilities (π_c)

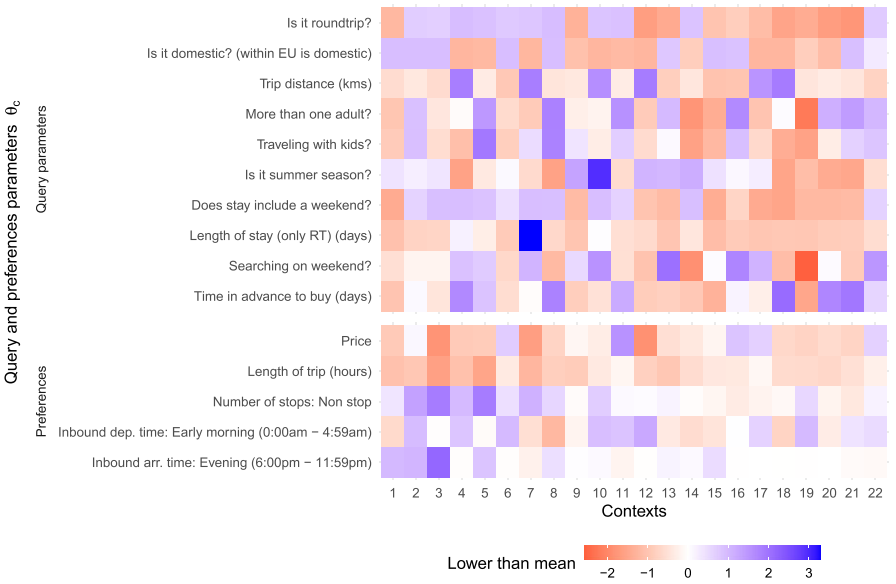
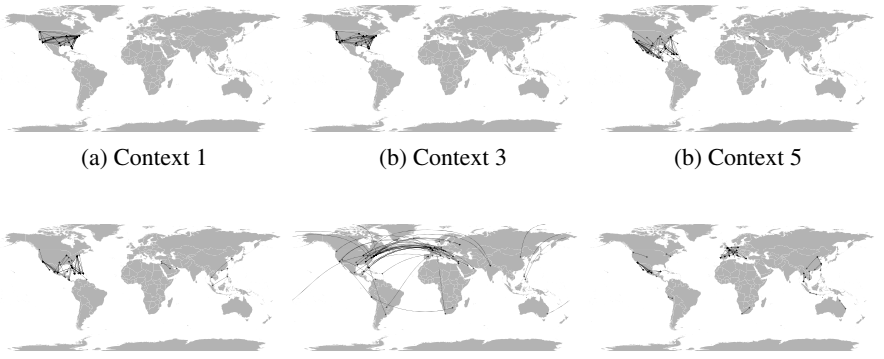


Fig. 8 Posterior mean of context location parameters, θ_c, relative to the average in the population. The top figure shows how each context deviates from the average with respect to the query variables. The bottom figure shows deviations with respect to the preference parameters. Blue (red) boxes mean positive (negative) deviation from the average in the population

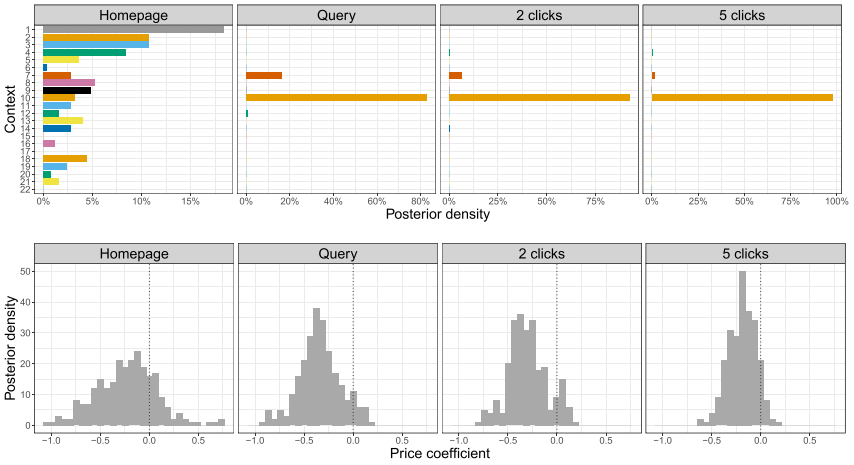
Similarly, we examine how the model infers price sensitivity as the customer moves across the journey (vertical histograms in the second rows of Fig. 10a and b). At homepage, both distributions are similar as both journeys belong to the same customer. After the query, both distributions tighten, with a lower price coefficient for *Journey*



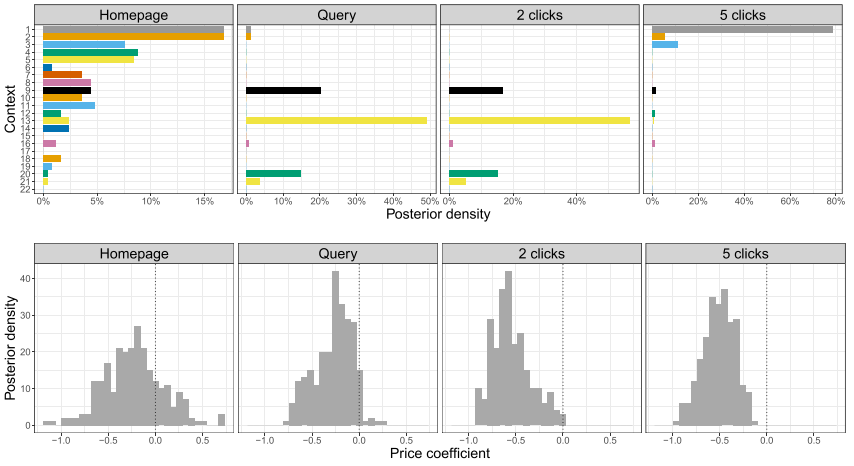
(a) Context 1 (b) Context 3 (b) Context 5

(d) Context 9 (e) Context 10 (f) Context 13

Fig. 9 Top 50 routes per context



(a) *Journey 1*: One adult roundtrip, August 11th-31st, from Kyiv, Ukraine (Boryspil International Airport) to New York (Newark Liberty International Airport), NY, USA.



(b) *Journey 2*: One adult oneway, September 12th, from Kyiv, Ukraine (Boryspil International Airport) to Lisbon, Portugal (Humberto Delgado Airport).

Fig. 10 Posterior context ($p(z_{ij}|\text{Data}_{ijt})$) and posterior price sensitivity ($p(\beta_{ij,\text{price}}|\text{Data}_{ijt})$) for two journeys of the same (sampled) customer at different stages of the journey, t: (1) homepage, (2) after query, (3) after 2 click steps, (4) after 5 click steps

2 than *Journey 1*. Interestingly, while both journeys belong to the same user, their inferred price sensitivity differs between journeys, especially after observing some clicks. In particular, after 5 clicks (steps), the model identifies that the customer is more price-sensitive for the flight to Lisbon than for the flight to New York. This is consistent with the fact that, across the population, journeys in context 1 exhibit stronger price sensitivity.

This illustrative example underscores the ability of the model to seamlessly integrate information from diverse customer interactions. By doing so, it extracts contexts

representing different customers' needs and generates insights into customers' preferences, empowering firms to continuously refine their understanding of what customers seek as they move along the journey.

4.4 Model predictive validity

While the model's primary objectives extend beyond predictions, confirming the model's precision in forecasting customer behavior serves as an indicator of its ability to harness and meaningfully leverage the data collected throughout the first-party journey. Specifically, we evaluate the models' ability to predict transactions and product choices at different points along the journey (e.g., immediately after a customer inserts the query, after two clicks, etc.). We compare these predictions with the predictions of conventional ML methods trained to predict those exact outcomes based on a large set of features that capture the customer information available at each point in the journey. We compare our model to the Random Forest (RF) and XGBoost models (see Web Appendix H for details of the prediction exercise).

As can be seen in Table 4, we find that our model performance is on par, and overall more robust across prediction tasks, relative to the ML models. Our modeling approach predicts reasonably well right after a query when no click information is available on the current journey, performing as well as the best alternative, which only does well either after query or after 5 steps. This result suggests our model can better encode information from past journeys and combine them with current journeys when data becomes available relative to traditional ML approaches (Smith et al., 2023). This relative performance is impressive because the ML models were specifically trained to predict each of the tasks (e.g., predict the focal outcome after a query or predict the focal outcome after 5 clicks/steps), whereas the proposed model was trained to piece together sources of information to customer needs and the context of the focal customer journey.

5 The value of first-party data

We first quantify the value of 1PD by comparing the predictive accuracy of the model insights at different stages of the customer journey and under different data reten-

Table 4 Predicting purchase incidence and product choice using the proposed and traditional machine learning models

	Incidence		Product choice given purchase			
	Balanced accuracy		Balanced accuracy		Hit rate	
Model	After query	After 5 clicks	After query	After 5 clicks	After query	After 5 clicks
Proposed model	0.62	0.65	0.58	0.81	0.16	0.62
Random forest	0.60	0.70	0.58	0.59	0.16	0.19
XGBoost	0.50	0.59	0.51	0.81	0.03	0.62

tion scenarios. Then, we illustrate two applications of how firms could leverage the model predictions from 1PD collected throughout the customer journey to make better marketing decisions.

5.1 The predictive value of first-party data

We quantify the value of 1PD at three levels: (1) the value of using the *current journey* data, (2) the value of using past *click-stream* data in addition to purchase history data, and (3) the value of tracking customers across visits. Before presenting our findings, we detail our approach, which uses both our full model and nested versions with subsets of 1PD to predict customer choices. To simulate unforeseen journeys, we use a set of held-out journeys, treating the characteristics of the chosen flight as the "ground truth" of what the customer was looking for. Predictions are made at three key "moments": when a query is inserted, after two clicks (steps), and after five clicks (steps).

Consistent with the analysis presented in Section 4.4, we compute the probability that the customer would select the (chosen) product based on the model's parameters at each point in time. We also explore the model's ability to predict the main attributes of the chosen flight (namely alliance, number of stops, price, and outbound length), which might represent a more realistic goal for the platform. We use hit rates for product choice, alliance and the number of stops (as they are categorical variables); and root mean squared error (RMSE) for price and length, which are continuous. Web Appendix I provides further details about predictions' calculations.

The value of data from the current journey We measure the value of the data collected along the *current journey* by comparing the performances of the `Full` model across different stages in the customer journey (the columns in Table 5). This analysis aims to highlight the benefits of incorporating customer information "live" as the customer progresses through the journey. We can see that the model significantly improves its accuracy of inferring what the customer is looking for as more data becomes available along the current journey. While this finding should not be surprising — as customers click, they often get closer to purchase — it is reassuring that the model is able to integrate such information in an effective manner. This result also highlights the premise of the paper of leveraging the customer journey as a source of information. Across all prediction tasks, the model's ability to predict what customers are looking for increases notably after observing a few clicks (e.g., the hit rate for the product chosen increases by 70% just after two clicks, and by nearly 300% after five clicks) relative to the query step. Exploring the main product attributes in this setting, we find that the (incremental) value of current journey clicks seems particularly strong when inferring what alliance the customer may choose, with hit rates increasing from 0.44 at query and 0.76 after five clicks.

The value of click-stream data To quantify the incremental value of prior-to-purchase data, we compare the `Full` model with a traditional hierarchical model of purchase, labeled `Only purchase`, which relies solely on purchase data and accounts for

Table 5 Model's accuracy at predicting the exact product chosen as well as the main attributes of the (chosen) product, evaluated at different stages of the journey

	After...		
	Query	2 clicks	5 clicks
Product choice (hit rate, ↑ better)			
Full model	0.16	0.27	0.62
% dif. vs. query		(+69.88%)	(+289.16%)
Only purchase	0.07	0.07	0.07
No customer tracking	0.10	0.29	0.62
Alliance (hit rate, ↑ better)			
Full model	0.44	0.55	0.76
% dif. vs. query		(+25.33%)	(+73.36%)
Only purchase	0.39	0.39	0.39
No customer tracking	0.32	0.46	0.71
Number of Stops (hit rate, ↑ better)			
Full model	0.59	0.62	0.78
% dif. vs. query		(+5.84%)	(+32.14%)
Only purchase	0.43	0.43	0.43
No customer tracking	0.39	0.50	0.72
Price (RMSE, ↓ better)			
Full model	0.96	0.92	0.81
% dif. vs. query		(-4.13%)	(-15.30%)
Only purchase	1.16	1.16	1.16
No customer tracking	1.15	1.07	0.91
Length (RMSE, ↓ better)			
Full model	0.88	0.84	0.73
% dif. vs. query		(-4.16%)	(-16.52%)
Only purchase	1.11	1.11	1.11
No customer tracking	1.20	1.10	0.93

Prediction measures: Hit rate (product, alliance, and number of stops) and RMSE (price and length). Performance of `Full` vs. `Only purchase` vs. `No customer tracking` models

customer heterogeneity ($\boldsymbol{\beta}_{ij} = \boldsymbol{\mu}_i$). This benchmark ignores past click information and fails to update during the current journey, similar to models commonly used for scanner panel data (e.g., Allenby & Rossi, 1998). Pre-purchase data is crucial because traditional models are limited when transactions are infrequent. In contrast, first-party data (1PD) collected throughout the customer journey-both current and past-offers a richer understanding of customer behavior, helping firms optimize marketing strategies to better meet customer needs. This highlights the importance of integrating the multiple behaviors observed along the journey, which are easy to collect but often not stored by companies.

The results in Table 5 suggest that there is significant value in the information collected prior to purchase throughout all stages of the customer journey. Specifically,

the hit rate for the product chosen after query is more than twice as large for the `Full` model (0.16) than for the `Only purchase` model (0.07). Additionally, the `Only purchase` model consistently predicts 15%-20% worse than the `Full` model for flight attributes after query. These findings indicate that queries and clicks from *prior journeys* allow the model to make better predictions even before leveraging the click information from the current journey. Looking across columns, we observe that the benefit of utilizing prior-to-purchase data is even more pronounced when incorporating query and clicks from the current journey. For example, the hit rate for the product chosen after five clicks is nine-fold higher than the model that leverages only purchase data.

The value of tracking customers Finally, to measure the value of tracking users, we contrast the performance of the `Full` model against a nested version labeled `No customer tracking`, where the customer identity is unknown. This (nested) model considers every journey as if it belongs to a "new" customer (for whom we don't know individual preferences) as journeys cannot be attributed to a specific customer,[16] and can either represent a situation of customers using the platform without logging in (and no cookie tracking) or a privacy scenario where the firm is not allowed to store customer historical data. This comparison helps us quantify the losses incurred when/if the company loses its ability to identify customers. Importantly, we compare both models at different moments of the journey, enabling us to measure the trade-off between the information lost due to the inability to track users and the information gained from the ability to collect and integrate data along the current journey.

Not surprisingly, not being able to track users hinders the model's ability to infer what they might be looking for (i.e., `No customer tracking` underperforms `Full` in most cases). The model particularly benefits from identifying a customer when making inferences early in the journey. Across all attributes, the full model performs substantially better than a model with no tracking information at query (e.g., a 38% increase on predicting Alliance). The disparity diminishes significantly as more clicks are observed in the current journey. After five clicks, even when the firm is unable to identify consumers, the model performs comparably to the model with complete information (62% hit rate in product choice). These results underscore the importance of incorporating customer interactions throughout the entire journey, particularly in cold-start scenarios where the company lacks prior information about the customer.

For individual product attributes, we find that leveraging past journeys enhances the model's ability to discern the type of product the customer seeks (e.g., for Alliance 76% hit rate for the `Full` model vs. 71% for the `No customer tracking` model; and for stops 78% hit rate for the `Full` model vs. 72% for the `No customer tracking` model). As we demonstrate next, individual attributes can be useful in a retargeting effort. For example, featuring specific brands (e.g., United flights from $199), product attributes (e.g., non-stop flights from $199), or a combination of these (e.g., United non-stop flights from $199).

[16] For this model, we keep the specification of the model, $\boldsymbol{\beta}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\rho}_j$, where the stable preferences term $\boldsymbol{\mu}_i$, in this case, captures journey-specific idiosyncratic preferences not explained by context. This allows the model to still learn, beyond context, from each click in the current journey.

Table 6 Product placement: The focal company uses the insights from the model to identify the (expected) preferred product and places it on top of the ranking

Model	P(abandon)	CTR (1st page)	CTR (Recommended)
Baseline: Cheapest / Observed Ranking	0.228	0.224	0.091
Our Recommended Product on Top	0.198	0.291	0.198

Taken together, these results underscore the value of collecting and integrating real-time 1PD, and combining those data sources across journeys, even when customers cannot be tracked. We find that the model's ability to leverage information along the current journey (combined with past journeys from "anonymous" customers) can compensate for the loss in firm's ability to identify the individual customer.

5.2 Leveraging model predictions for managerial decisions

This section outlines the practical significance of our modeling framework by examining its application to two prevalent marketing decisions in contexts where firms can collect 1PD throughout the customer journey: (1) product placement (or recommendation), and (2) retargeting. Although direct empirical validation of the model is beyond our scope, we utilize holdout data to simulate potential marketing outcomes, comparing the use of first-party customer journey data against baseline methods.

5.2.1 Product placement / recommendation

We explore a scenario where, right after the customer inserts the query, the platform wishes to place at the top of the ranking the product the platform believes the customer would prefer the most based on its current understanding of user preferences at that stage of the journey. This is similar to the "recommended" or "best" product commonly presented at the top of the ranking in travel platforms.[17] Using holdout journey data, we compare the standard practice (baseline) of presenting products in the observed sequence, typically arranged by price, with an alternative scenario. In this alternative, the firm highlights at the top of the ranking the product most likely to match consumer preferences based on our model, followed by the original ranking from the data (the average position of the recommended product in the original ranking is 15.9).

We simulate customer responses using our proposed model[18] and report (1) the likelihood of the customer abandoning the journey after seeing the product results, (2) the chance of clicking on any product from the initial page, and (3) the probability of the customer clicking the top product in the ranking (lowest priced product in the observed data and the recommended product in our counterfactual).

[17] The platform we worked with did not have such a placement during our data period.

[18] Our reliance on model-generated evaluation is necessary because we do not observe the counterfactual scenarios in the holdout data (Smith et al., 2023). We generate such evaluations using the full model with all click and purchase information, which has different estimates from those used only using search query.

The scenario corresponding to the baseline condition, where products were sorted by price, results in a customer journey abandonment probability of 22.8%, a click-through rate (CTR) of 22.4% on any product on the first page, and a mere 9.1% CTR for the top, often lowest price, product. In contrast, leveraging insights from our model substantially and significantly reduces the abandonment rate to 19.8%, increases the first-page CTR to 29.1%, and boosts the recommended product CTR to 19.8%. (see Table 6).

5.2.2 Retargeting

Retargeting is another prevalent marketing practice aimed at encouraging purchases among customers who have shown interest in a purchase but have not completed the journey. For instance, a company might send an email or show a display ad featuring a flight that aligns with the customer's previous searches. This is a common practice in the industry (Bloomreach, 2022; Google, 2023). In this application, we posit that the company can capitalize on the information gleaned from the customer journey to determine the creative for the retargeted advertisement. We propose to present in the retargeting creative the product deemed most likely to be purchased by the customer, as predicted by our model. We contrast this approach with two alternative heuristics: (1) the most popular product, that is, the product with the highest non-personalized predicted purchase probability in our dataset, and (2) the last product the customer clicked on.[19] Specifically, we simulate outcomes for a single ad impression of a retargeted ad for each customer where we vary the product featured in the impression, measuring the click-through-rates on these ads.[20] To simulate the behavior, we use the full model with all click and purchase information (which is different from the model used to set the "best product" policy), where we assume the utility of clicking on a retargeted ad is proportional to the utility estimated in the model.[21]

The results, depicted in Table 7, illustrate the click-through rates (CTRs) for retargeted ads based on the proposed model and the two baseline alternatives. The first baseline model based on the most popular product has a CTR of 5.4%, whereas the baseline model, which retargets based on the last clicked item, achieves a CTR of 5.6%. In contrast, retargeting the best product based on our model leads to a superior CTR of 7.4%. This represents a CTR improvement of 28%-32% relative to retargeting using the two baseline cases.

Collectively, these findings emphasize the practical benefit of the proposed model for businesses looking to increase customer engagement and sales conversions. By leveraging the insights from the proposed model, companies can not only improve

[19] When no product has been clicked on, we feature the most popular product in our dataset for that query.

[20] Our analysis includes customer journeys with a minimum of two clicks, applying the inferences of the model at that specific juncture to find the best product.

[21] Since the variance of the error term for the retargeting action is not known, we set it to $\sigma = 1.25$ and vary it in the Web Appendix J. We find that the improvements in CTRs to be robust to alternative standard deviations. The intercept is set so that the mean CTR for a retargeted ad featuring a random product equals 2% across journeys (the average retargeting CTR without content personalization reported in the industry is 0.7% Signifi Media 2020, and highly personalized ads have 3x higher CTRs compared to non-personalized ads Bleier and Eisenbeiss 2015). For details, see Web Appendix J.

Table 7 Retargeting: Predicted CTR when the firm utilizes insights from the model to formulate a retargeting offer

Model	Click-through rates
Baseline 1 (Most popular)	0.054
Baseline 2 (Last clicked)	0.056
% dif. vs. most popular	(+3.05%)
Highest preference based on our model	0.074
% dif. vs. most popular	(+31.94%)
% dif. vs. last clicked	(+28.04%)

The baseline models presume that the retargeting offer features the product that was most recently clicked or the most popular product

product recommendations by effectively positioning products in a way that better aligns with the customer's preferences but can also enhance engagement by tailoring retargeted advertisements. This analysis demonstrates the potential of using our model to create more personalized and impactful customer interactions. This can lead to a more efficient allocation of marketing resources and potentially higher returns on investment, making the proposed model a valuable tool for data-driven marketing efforts.

6 Conclusion

We propose a probabilistic non-parametric Bayesian machine learning model that integrates information collected along the first-party journey and combines such information across customers and journeys. This model uses within-journey data to identify distinct journey-specific preferences. Information from past journeys, including purchases, searches, and clicks, helps infer stable customer preferences. Behavior across customers enriches information by integrating data from customers with analogous context-specific preferences. This approach not only helps firms to infer what customers are looking for, even in settings where purchases occur infrequently, but also addresses the so-called "cold start problem," offering a solution to compensate for the lack of historical data with real-time data collected along the first-party customer journey.

Applying the model to data from an online travel platform showcases its ability to extract key aspects of customer preferences, such as airline choices, flight stop preferences, and price considerations. The model demonstrates significant adaptability, by continuously updating inferences as customers progress throughout their journeys. This is a notable improvement over traditional panel-data models that rely solely on purchase data. Our model predicts consumers' product selection nine times better than models that only use purchase or sales data, highlighting the importance of pre-purchase data. Furthermore, our modeling framework assesses the implications of data loss under heightened data privacy scenarios, illustrating how leveraging infor-

mation across first-party customer journeys can compensate for reduced customer identification capabilities. This is increasingly relevant as platforms allow searches without requiring login credentials or are mandated to limit the storage of customers' data.

Our research has several limitations that offer promising directions for future exploration. First, the complexity of our model may challenge computational efficiency and scalability, especially with large datasets. Faster approximation strategies for the posterior distribution of customer preferences, such as point estimates or variational approximations, may be necessary for real-time applications. Modern inference methods, like amortized variational inference (Kingma & Welling, 2013; Agrawal & Domke, 2021), could also be used to efficiently update journey-specific preferences. Second, while we emphasize the benefits of incorporating first-party data (1PD), our model does not account for external factors that influence customer behavior, such as economic trends or unexpected events. This could limit the model's ability to leverage past journeys to augment limited early data in the customer journey. Future research should explore incorporating broader dynamics or shocks to allow the model to adapt to significant shifts in preferences.

Third, our model is not based on a rational economic framework, which prevents it from deriving structural invariant parameters useful for counterfactual applications that could alter market equilibrium. Future work should aim to integrate the flexibility of our approach with structural search models. Fourth, due to infrequent purchase behavior in our data, we assume a common context distribution across customers. In settings with more frequent journeys, sufficient histories could allow for customer-specific contexts. Alternative empirical settings where journeys are observed across multiple product categories may permit combining information across journeys, customers, and product categories.

Finally, we measure the value of within-journey and past-journey data for an online travel platform, but this value may vary by application. For example, within-journey data may be more valuable for durable goods than for past journeys from years ago. Investigating the varying value of customer journey data across different contexts could be a fruitful area for future research. We hope our work inspires further studies on leveraging first-party customer journeys and their implications for marketing strategies.

In conclusion, we aim to highlight the opportunity to fuse historical purchase and click-stream data with current first-party journey data in customer relationship management and choice modeling research. This perspective is particularly relevant given the evolving challenges posed by data privacy concerns and the dynamic and non-linear nature of customer journeys.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11129-024-09287-y.

References

Agrawal, A., & Domke, J. (2021). Amortized variational inference for simple hierarchical models. *Advances in Neural Information Processing Systems, 34*, 21388–21399.

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

Ansari, A., & Mela, C. F. (2003). E-Customization. *Journal of Marketing Research, 40*(2), 131–145.

Atchadé, Y. F., & Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli, 11*(5), 815–828.

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

Bleier, A., & Eisenbeiss, M. (2015). Personalized online advertising effectiveness: The interplay of what, when, and where. *Marketing Science, 34*(5), 669–688.

Bloomreach (2022). Enhance your personalization with real-time retargeting. https://www.bloomreach.com/en/blog/2022/real-time-targeting/. Accessed: 2024-06-08.

Boughanmi, K., & Ansari, A. (2021). Dynamics of musical success: A machine learning approach for multimedia data fusion. *Journal of Marketing Research, 58*(6), 1034–1057.

Braun, M., & Bonfrer, A. (2011). Scalable inference of customer similarities from interactions data using Dirichlet Processes. *Marketing Science, 30*(3), 513–531.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In 2010 20th international conference on pattern recognition, pp. 3121–3124. IEEE.

Bronnenberg, B. J., Kim, J. B., & Mela, C. F. (2016). Zooming in on choice: How do consumers search for cameras online? *Marketing Science, 35*(5), 693–712.

Bruce, N. I. (2019). Bayesian nonparametric dynamic methods: Applications to linear and nonlinear advertising models. *Journal of Marketing Research, 56*(2), 211–229.

Chen, Y., & Yao, S. (2017). Sequential search with refinement: Model and application with click-stream data. *Management Science, 63*(12), 4345–4365.

Dew, R., Ansari, A., & Li, Y. (2020). Modeling dynamic heterogeneity using gaussian processes. *Journal of Marketing Research, 57*(1), 55–77.

Dew, R., Padilla, N., Luo, L. E., Oblander, S., Ansari, A., Boughanmi, K., Braun, M., Feinberg, F. M., Liu, J., Otter, T., et al. (2024). Probabilistic machine learning: New frontiers for modeling consumers and their choices. Available at SSRN 4790799.

Donnelly, R., Kanodia, A., & Morozov, I. (2024). Welfare effects of personalized rankings. *Marketing Science, 43*(1): 92–113.

Feit, E. M., Wang, P., Bradlow, E. T., & Fader, P. S. (2013). Fusing aggregate and disaggregate data with an application to multiplatform media consumption. *Journal of Marketing Research, 50*(3), 348–364.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In Computing science and statistics: Proceedings of the 23rd symposium on the interface, volume 571, page 578. Fairfax, Virginia: Interface Foundation of North America, Inc.

Google (2023). How jumia used first-party data to boost app remarketing effectiveness. https://www.thinkwithgoogle.com/intl/en-emea/marketing-strategies/data-and-measurement/jumia-first-party-data-remarketing/. Accessed: 2024-06-08.

Grewal, D., & Roggeveen, A. L. (2020). Understanding retail experiences and customer journey management. *Journal of Retailing, 96*(1), 3–8.

Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2016). Session-based recommendations with recurrent neural networks. arXiv preprint arXiv:1511.06939.

Honka, E., & Chintagunta, P. (2017). Simultaneous or sequential? Search strategies in the U.S. auto insurance industry. *Marketing Science, 36*(1), 21–42.
Honka, E., Seiler, S., & Ursu, R. (2024). Consumer search: What can we learn from pre-purchase data? *Journal of Retailing, 100*(1), 114–129.
Huberman, E. (2021). First-party data collection is more crucial than ever. *Forbes*, December 21.
Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association, 96*(453), 161–173.
Iyengar, R., Ansari, A., & Gupta, S. (2003). Leveraging information across categories. *Quantitative Marketing and Economics, 1*(4), 425–465.
Jacobs, B. J., Donkers, B., & Fok, D. (2016). Model-based purchase predictions for large assortments. *Marketing Science, 35*(3), 389–404.
Kim, J. B., Albuquerque, P., & Bronnenberg, B. J. (2010). Online demand under limited consumer search. *Marketing Science, 29*(6), 1001–1023.
Kim, J. G., Menzefricke, U., & Feinberg, F. M. (2004). Assessing heterogeneity in discrete choice models using a Dirichlet Process prior. Review of Marketing Science.
Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114.
Koulayev, S. (2014). Search for differentiated products: Identification and estimation. *RAND Journal of Economics, 45*(3), 553–575.
Lee, L., Inman, J. J., Argo, J. J., Böttger, T., Dholakia, U., Gilbride, T., Van Ittersum, K., Kahn, B., Kalra, A., Lehmann, D. R., et al. (2018). From browsing to buying and beyond: The needs-adaptive shopper journey model. *Journal of the Association for Consumer Research, 3*(3), 277–293.
Liu, J., & Toubia, O. (2018). A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science, 37*(6), 930–952.
Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Marketing Science, 23*(4), 579–595.
Morozov, I. (2023). Measuring benefits from new products in markets with information frictions. *Management Science, 69*(11), 6988–7008.
Morozov, I., Seiler, S., Dong, X., & Hou, L. (2021). Estimation of preference heterogeneity in markets with costly search. *Marketing Science, 40*(5), 871–899.
Murphy, K. (2022). With first-party data, marketers are finally in the driver's seat. *Harvard Business Review Online*, September 27.
Neal, R. M. (2000). Markov chain sampling methods for Dirichlet Process mixture models. *Journal of Computational and Graphical Statistics, 9*(2), 249–265.
Padilla, N., & Ascarza, E. (2021). Overcoming the cold start problem of customer relationship management using a probabilistic machine learning approach. *Journal of Marketing Research, 58*(5), 981–1006.
Pitman, J., & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability, 25*(2), 855–900.
Rossi, P. E., McCulloch, R. E., & Allenby, G. M. (1996). The value of purchase history data in target marketing. *Marketing Science, 15*(4), 321–340.
Sarwar, B., Karypis, G., Konstan, J., & Reidl, J. (2001). Item-based collaborative filtering recommendation algorithms. In Proceedings of the tenth international conference on World Wide Web - WWW '01, volume 3, pages 285–295, New York, New York, USA. ACM Press.
Seiler, S. (2013). The impact of search costs on consumer behavior: A dynamic approach. *Quantitative Marketing and Economics, 11*(2), 155–203.
Signifi Media (2020). 8 remarketing statistics you need to be aware of in 2020. https://www.signifimedia.com.au/8-remarketing-statistics-you-need-to-be-aware-of-in-2020/. Accessed: 2024-06-08.
Smith, A. N., Seiler, S., & Aggarwal, I. (2023). Optimal price targeting. *Marketing Science, 42*(3), 476–499.
Ursu, R. M. (2018). The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science, 37*(4), 530–552.
Weitzman, M. L. (1979). Optimal search for the best alternative. *Econometrica, 47*(3), 641–654.
Yoganarasimhan, H. (2020). Search personalization using machine learning. *Management Science, 66*(3), 1045–1070.