

Clickstream Data and Inventory Management: Model and Empirical Analysis

Tingliang Huang

Department of Management Science & Innovation, University College London, t.huang@ucl.ac.uk

Jan A. Van Mieghem

Kellogg School of Management, Northwestern University, vanmieghem@kellogg.northwestern.edu

We consider firms that feature their products on the Internet but take orders offline. Click and order data are disjoint on such non-transactional websites and their matching is error-prone. Yet, their time separation may allow the firm to react and improve its tactical planning. We introduce a dynamic decision support model that augments the classic inventory planning model with additional clickstream state variables. Using a novel data set of matched online clickstream and offline purchasing data, we identify statistically significant clickstream variables and empirically investigate the value of clickstream tracking on non-transactional websites to improve inventory management.

We show that the noisy clickstream data is statistically significant to predict the propensity, amount, and timing of offline orders. A counterfactual analysis shows that using the demand information extracted from the clickstream data would reduce the inventory holding and backordering cost by about 5% in our data set. Finally, robustness checks of incorporating more customer heterogeneity and dynamic behavior yield interesting findings as a byproduct: Existing customers are forward-looking in their information collection, and are more loyal than new customers.

Key words: Inventory Theory and Control, Dynamic Programming, Econometric Analysis, Structural Estimation, Empirical Research, OM-Information Technology Interface

1. Introduction and Related Literature

Recent Internet clickstream tracking technology has generated the fast growing practice of web analytics and extensive ongoing research in academia. Indeed, the Internet has changed the way business works by providing new information and distribution channels for both firms and customers. Customers can readily obtain product information online without physically visiting a firm. Firms can use clickstream tracking technology to see in real time who is visiting their websites and analyze detailed clickstreams to learn more information in advance.

Clickstream tracking allows firms to “learn about customers without asking” (Montgomery and Srinivasan 2003), but the associated academic research has been largely focused on online shopping and e-commerce: Montgomery (2001) shows that quantitative models that are commonly used in brick-and-mortar distribution channels prove to be useful in optimizing the use of clickstream data. The associated literature is extensive, see e.g., Johnson et al. (2003), Moe and Fader (2004),

Montgomery et al. (2004), Sismeiro and Bucklin (2004), Van den Poel and Buckinx (2005), Hui et al. (2009) and references therein. This literature is essentially about the marketing benefits of clickstream tracking because e-commerce websites serve primarily as sales channels. Clickstream tracking allows e-commerce firms to get accurate readings of the efficiency of their websites, quickly usher a visitor (referred to as “she” throughout the paper) who is about to purchase an item to a high-speed server, identify target visitors to show pop-up coupons, and so on.

In contrast to e-commerce settings, we investigate “non-transactional websites” that serve predominantly as a product catalog while orders are taken offline. Many B2B settings as well as some B2C settings fall in this category. Specifically, this study stems from our interaction with a U.S. manufacturer of industrial products, hereafter referred to as “the company.” The company makes high-end roll-up doors that are customized for industrial and commercial buildings with regards to size, type of material, type of environment, etc. The doors can go into new buildings or can replace older doors. Prices for a door range from the thousands to tens of thousands of dollars. Like many others, the company provides current and potential customers with company, product, and contact information on its website. In contrast to e-commerce firms, however, the website is non-transactional and the company sells its products offline, either direct or through dealers. The company hires the services of a web analytics firm that specializes in clickstream tracking to help demand forecasting, procurement and inventory planning.

Our study focuses on the operational benefits of clickstream tracking by investigating its use as advance demand information for procurement, production and inventory planning. We are interested in how, and to what extent, clickstream data from non-transactional websites can improve demand forecasting for inventory management. The clickstream data and sales data we study has significant differences from the data from e-commerce stores studied in the literature because the company website is non-transactional and the customers are also firms. These differences result in both challenges and opportunities:

There are two challenges of non-transactional websites. First and foremost, while it has been confirmed in the literature that online click behavior should be correlated with purchasing behavior in e-commerce settings, it is much less clear whether such correlation persists in non-transactional settings because customers do not have to visit the website to make a purchase. This procedural separation reduces the predictive power of web visits to forecast purchase orders if there is any statistical relationship between them at all.

A second challenge concerns the identification of web visitors. In our setting web visitors do not identify themselves because they do not purchase and reveal contact or payment information online.

The firm can only learn each visitor’s identity through her IP address. In addition, we study a business-to-business (B2B) setting where the customers themselves are firms. This has benefits and costs: About 82% of the visits in our clickstream data come from a company-registered IP address so that the visitor is easily identified with a company. Then we can manually match clickstream data with sales data to investigate the correlation between clicking behavior and ordering behavior. The other 18% of visits come from large service provider IP addresses (e.g.,@comcast.com, @cox.com, @att.com)—perhaps visits from home-computers or cellular devices, which prevents the identification of the visitor and the matching with order data. These visits are deleted from the data set. While one expects corporate online browsing behavior to be less frivolous than that in a business-to-consumer (B2C) setting, another challenge is that we cannot identify the various individuals who are involved in the purchasing process. Only IP addresses are tracked, typically at the level of a firm’s computer center/connection to the Internet but not at the level of individual computers inside the firm. Therefore, the unit of analysis in our data is a firm and all visits from a firm (even if the firm has multiple locations) are aggregated and indistinguishable from one visitor. In addition, a potential customer may also browse the website from her home computer(s). Thus, our clickstream-order data is more noisy than in e-commerce.

Meanwhile, non-transactional websites provide the opportunity for firms to react. The longer time separation between clicks and orders also has an important benefit: If it exceeds the production or procurement lead time, the firm can respond to changes in advance demand information. Indeed, the company we study observes lead times on the order of weeks and even months. It then can accrue operational benefit by better matching supply with forecasted demand and reducing its inventory holding and backordering cost. (In contrast, the operational value of clickstreams may be trivially small in the e-commerce setting like Amazon, because the time lag between clicks and orders can be on the order of minutes, too short to adjust operational plans.) For this reason, operations managers as well as sales persons could benefit significantly by studying clickstream data in our setting.

In this setting of a B2B business with non-transactional informational websites, we address the following research questions: (1) How can we use clickstream data in inventory management? This requires a tactical model that explicitly incorporates clickstream data in operations management. (2) How can we identify the significant clickstream data and prediction functions (needed in the model) and improve the demand forecast? (3) How large is the operational value of using the advance demand information from clickstreams to reduce inventory holding and backordering costs in our setting?

We believe these questions are timely and important for the following reasons.

First, the recent fast-growing research using clickstream data has already demonstrated the great interest and importance for e-commerce firms. The same applies to offline-selling firms. Second, the web analytics industry is also fast growing. Understanding consumer online browsing behavior and its value helps firms make investment decisions regarding the adoption of clickstream tracking technology.

Last and most importantly, these questions are of great relevance from an operations management perspective. Matching supply with demand is one of the main issues for operations management. There is a vast body of literature modeling advance demand information; see, for example, Hariharan and Zipkin (1995), Raman and Fisher (1996), Chen (2001), Gallego and Özer (2001, 2003), Özer and Wei (2004), Tan et al. (2007), Wang and Toktay (2008), and Gayon et al. (2009). Özer (2011) provides a comprehensive literature review. All these papers assume that advance demand information is available and study how to use it in inventory management. On one hand, our study is in the same spirit of and complementary to this literature by introducing a practical decision support model that endows classic inventory management with clickstreams as a flow of advance demand information. On the other hand, our study is the logical precedent: To what extent can advance demand information be obtained from clickstreams? Although the value of advance demand information is well established and understood theoretically, research on how advance demand information is obtained in practice and its empirical evidence seems largely absent in the operations management literature. Özer (2011) offers several examples of obtaining advance demand information in practice such as flexible delivery at the time of ordering, ordering customized products, and advance selling. All these practices share the same feature that advance demand information is obtained at the time of customer ordering. Clickstream data, in contrast, provides advance demand information in a completely different way: First, it can be unrelated to customer ordering. Second, such information can be obtained well before customer ordering. (For example, the earliest lead time in our data set is 438 days before a customer actually placed an order and the mean time is around 90 days.) Hence, this kind of demand information can be truly “advance.” More importantly, such information is obtained “without asking” customers, which is also called “inferring” (Fay et al. 2009). Our empirical study of this novel information technology shows that clickstream data is useful for operation managers to predict demand and helps firms “do the right thing at right time in right quantities.”

The main contributions and findings of the paper are as follows:

1. We introduce a practical dynamic decision support model that augments the traditional inventory management with clickstreams as additional state variables in the dynamic programming formulation for demand forecasting.

2. We conduct an empirical study to identify *i*) which clickstream variables are statistically significant for demand forecasting, *ii*) how to include them into the state variables of the dynamic model, and *iii*) to estimate the extent to which utilizing the clickstreams generates operational value. We find that customer clicking behavior is a statistically significant predictor of the corresponding offline purchasing behavior, in terms of not only ordering probabilities and ordering amount (in monetary value), but also ordering timing (lead time). We show that using the information extracted from the clickstream data would reduce the inventory holding and backordering cost by about 5%.

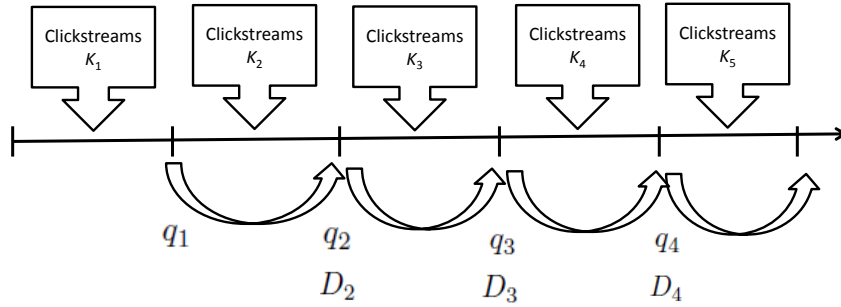
3. Our robustness checks yield several novel findings about customer behavior. For example, based on a structural estimation of a novel economic model of customer information collection, we show that, interestingly and surprisingly, existing customers are forward-looking in their information collection and are more loyal, i.e., more likely to return to the website than new customers. We treat these findings as a byproduct of our study.

The outline of this paper is as follows. The next section presents a theoretical model to demonstrate how clickstream data can be used to improve demand forecasting and inventory management. In §3, we empirically identify the clickstream variables that are significant for demand forecasting. In §4, we quantify the operational value of advance demand information from the clickstream data using our model. In §5, we carry out robustness checks by considering more customer heterogeneity and dynamic customer behavior. §6 contains the discussions and limitations.

2. A Model of Using Clickstream Data in Inventory Management

We start by introducing a tactical model of using clickstream data in demand forecasting and inventory management that can serve as a decision support system in practice. This practical model endows classic inventory management with clickstreams as a dynamic flow of advance demand information. In §3, we will empirically identify relevant model variables. This model will also be our tool for estimating the operational value of clickstream data in §4.

We explain how to use clickstreams in inventory management first in a single-period newsvendor model, and then in a multi-period dynamic model. In a single-period model, before the company's production or procurement decisions, clickstreams are observed to predict demand. For each visitor i who clicked, the company can use clickstreams to estimate her purchasing probability $P_i \equiv f(X_i)$

Figure 1 Description of the Dynamic Programming Model

for $i = 1, 2, \dots$, where X_i is a vector of independent variables including clickstream variables and f denotes a general prediction function. We shall empirically specify both X_i and f in the next section. Assuming all the visitors are independent decision makers, a simple combinatorial calculation then yields the predicted distribution of the total demand that can be used to derive the optimal inventory for this single-period newsvendor model.

To explain how to use clickstream data in a dynamic setting, consider a discrete-time inventory control model endowed with clickstream data. Suppose there are T replenishment periods. In each period t , the company can observe the clickstreams for each visitor i who clicked in this period. To formulate the company's inventory control problem as a dynamic programming problem, we need a description of the company's operations. In each replenishment period t , the company first satisfies or backorders any realized demand D_t and observes clickstreams of new visitors $\mathbf{K}_t = (K_{1,t}, K_{2,t}, \dots, K_{J,t})$. Then it updates its demand forecast, and determines its ordering quantity q_t . This cycle repeats, as depicted in Figure 1.

Extending the previous single-period model to a multi-period model introduces significant analytical complications for at least three reasons: First, the demand distribution in period t depends on what happened in previous periods. Second, visitors are heterogeneous. Third, different from the single-period model where a customer either "purchases" or "never purchases," she has an additional decision: wait and may purchase later. The model has to keep track of the richness of the system dynamics. We adopt the following approach: (i) To account for visitor heterogeneity while still retaining analytical tractability, we classify all visitors into J classes or categories. Within each class j , each visitor is homogeneous, i.e., each visitor in class j who clicked in period t but had not clicked before has prior purchasing probability $p_{j,t'}$ in period t' for $j = 1, 2, \dots, J$ and $t' = t + 1, t + 2, \dots$. Choosing J is at the company's disposal. Intuitively, it is natural to assume that visitors who share the same value of the predictors X_i constitute a class. Similar to the single-period

model, the purchasing probability $p_{j,t'}$ can be estimated using $P_i \equiv f(X_i)$. The only difference is that we will use the empirical distribution of the click lead time to predict when (i.e., in which period) a purchase will occur. (ii) We assume that each visitor in class j has the prior probability $\gamma_{j,t'}$ in period t' for $t' = t + 1, t + 2, \dots$ of never purchasing the product. Clearly, $1 - p_{j,t'} - \gamma_{j,t'} \leq 1$, where the equality always holds in the single-period model but not necessarily in this model for the third reason we pointed out. Estimating these probabilities for *non-buyers* is trivial for the single-period model given the equality relationship, but can be difficult in the multi-period model. We will demonstrate how to estimate them in §4.

We are now ready to describe the system dynamics analytically. Our approach allows for a class-by-class analysis. We denote $K_{j,t}$ as *the number of new visitors of class j in period t* , meaning the visitors of class j who visited the website in period t but never visited the website *before* period t . This definition precludes the “double counting” as will become clear in the flow equation. Notice that we count “visitors” rather than “clicks” given that a visitor typically clicks multiple times. We will call these $K_{j,t}$ visitors *potential buyers*. They represent potential future demand, as they *may* convert to *real buyers* in future periods. For analytical convenience, we assume that each visitor buys at most one unit of the product. This assumption is reasonable in our setting of a durable industrial product. Let the random variable $Z_{j,t+1}$ denote the *total* number of potential buyers of class j at the beginning of period $t + 1$, i.e., the *cumulative* number of customers of class j who clicked up to period $t + 1$ and are still part of the potential buyers for future periods, i.e., they have not purchased or have not been identified as non-buyers yet. Then we have the dynamic flow equation as follows:

$$Z_{j,t+1}(z_{j,t}; \mathbf{H}_{j,t}) = z_{j,t} + K_{j,t+1} - D_{j,t+1}(z_{j,t}; \mathbf{H}_{j,t}) - L_{j,t+1}(z_{j,t}; \mathbf{H}_{j,t}), \quad (1)$$

which is the previous realized number $z_{j,t}$, plus the number of new potential buyers $K_{j,t+1}$ from the clickstreams observed in period $t + 1$, minus the demand $D_{j,t+1}$ and non-buyers $L_{j,t+1}$.¹ Notice that the terms in lower case denote the realizations of the random variables in upper case. In general, $Z_{j,t+1}$ depends on the entire “history” $\mathbf{H}_{j,t} \equiv (k_{j,1}, \dots, k_{j,t}; d_{j,1}, \dots, d_{j,t}; l_{j,1}, \dots, l_{j,t})$. Let $\mathbf{Z}_t \equiv (Z_{1,t}, Z_{2,t}, Z_{3,t}, \dots, Z_{J,t})$ and $\mathbf{H}_t \equiv (\mathbf{H}_{1,t}, \mathbf{H}_{2,t}, \mathbf{H}_{3,t}, \mathbf{H}_{J,t})$, then the state vector $(\mathbf{Z}_t, \mathbf{H}_t)$ completely describes the system in period t . Clearly, the total demand in period $t + 1$: $D_{t+1} = \sum_{j=1}^J D_{j,t+1}$.

¹ The non-buyers may not be observable from clickstreams, in which case $L_{j,t+1} = 0$. Typically companies can estimate the probability that a customer never purchases from clickstreams. Non-buyers can be easily observed in cases where the company can obtain some offline information by communicating with the visitors.

According to flow equation (1), \mathbf{Z}_{t+1} depends on the complete history \mathbf{H}_t . Working with this general non-Markovian model is analytically challenging. From now on, we will work with a Markovian model by assuming that all $z_{j,t}$ potential buyers have the same purchasing probability $p_{j,1}$ and never-purchasing probability $\gamma_{j,1}$ for any period $t \geq 1$ *given that* they did not purchase in previous periods. This assumption implies that $p_{j,t} = p_{j,1}(1 - p_{j,1} - \gamma_{j,1})^{t-1}$ and $\gamma_{j,t} = \gamma_{j,1}(1 - p_{j,1} - \gamma_{j,1})^{t-1}$ for $t \geq 2$. Hence, we can drop the dependence on \mathbf{H}_t , and the vector \mathbf{Z}_t suffices to fully describe the system in period t .

The Markovian assumption allows us to formulate the company's inventory management problem as a finite-horizon discounted dynamic programming problem using x , the inventory position, and \mathbf{z} , the *vector* of the cumulative number of potential buyers in each visitor-class over the future. Let $V_t(x, \mathbf{z}_t)$ denote the minimum expected discounted cost at state (x, \mathbf{z}_t) starting from the beginning of period t to the end of the planning horizon. We assume that any remaining inventory is salvaged with per-unit revenue that equals to the per-unit procurement cost c and any outstanding backorders are satisfied with per unit cost of c at the end of the planning horizon. Then we have

$$V_{T+1}(x, \mathbf{z}_{T+1}) = -cx.$$

For $t = 1, 2, \dots, T$, we have the Bellman equation:

$$V_t(x, \mathbf{z}_t) = \min_{y \geq x} \left\{ \underbrace{c(y-x)}_{\text{procurement cost}} + \beta_0 \left[\underbrace{h\mathbb{E}(y - D_{t+1}(\mathbf{z}_t))^+}_{\text{holding cost}} + \underbrace{b\mathbb{E}(D_{t+1}(\mathbf{z}_t) - y)^+}_{\text{backorder cost}} \right] + \beta_0 \mathbb{E} \left[\underbrace{V_{t+1}(y - D_{t+1}(\mathbf{z}_t), \mathbf{Z}_{t+1}(\mathbf{z}_t))}_{\text{optimal cost-to-go}} \right] \right\},$$

where β_0 is the usual time discount factor and y is the order-up-to level as the company's decision variable. This formulation is motivated by Gallego and Özer (2001), Özer (2011) and references therein where they include an observed part of lead time demand in classic inventory models (cf. Porteus 1990). While our inventory model endowed with clickstreams is novel, the dynamic flow of advance demand information \mathbf{z} extracted from these clickstreams \mathbf{K} essentially provides observable lead time demand in spirit. Using a similar technique as Gallego and Özer (2001), one can prove that the optimal inventory policy is "clickstreams-dependent" base stock policy. All the parameters required to evaluate the cost saving due to using clickstream data in §4 will be estimated from the data in our subsequent empirical study.

3. Empirical Analysis

In this section, we will empirically demonstrate that clickstreams are indeed useful to estimate the purchasing probability $P_i \equiv f(X_i)$ for $i = 1, 2, \dots$ in our model in §2. To this end, we first discuss our data sets and variable definitions, then specify the general prediction function f as a simple logit regression equation, and finally show which click variables among X_i are statistically significant.

3.1. Data Source and Characteristics

We use two data sets of the company that sells high-end roll-up doors in North America. The first data set is the clickstream data from August 26, 2006 to February 28, 2008, and the second is the sales data from August 26, 2006 on. There are 5185 customers, and 9694 visits in the data.

In the the clickstream data, the unit of data corresponds to a customer who clicked and has the following fields: the name of the customer identified from her IP address, the clickstream which is the recorded click behavior that includes the time of visits/clicks, visit frequency (i.e., the cumulative number of visits), average time stayed online per visit, average number of pages visited per visit, and the detailed page-specific data such as the sequences of pages visited and the time length.

Each unit in the sales data records the customer name, the ordering amount (in U.S. dollars), and the time of ordering.

Before statistical analysis could be commenced, several preprocessing tasks were executed. First, we cleaned the clickstream data by deleting unidentifiable clicks. For example, visitors who visited the website at home or from private computers show up with the unidentifiable IP address of their service provider. Processing these observations/visits deleted approximately 18% of the total visits in our data set.

The second preprocessing step deleted some organizations that we excluded in our study such as universities, public organizations, etc. In the ordering data set, indeed, no universities or public organizations ever purchased any product from the company. Their visits may have been research-inspired.

Third, as discussed in the introduction, we aggregated all the visitors within a company as a *single* visitor by their company names. The reason of doing this is simply because of the limitation in our information availability, i.e., the clickstream data only shows the company names, not the persons who actually visit.

Finally, we matched the clickstream data set with the sales data set together by the firm/customer names. We have 9694 visits in our clickstream data set after preprocessing and matching with the sales data.²

² Admittedly, this matching of clicks with orders could be noisy if individual companies have high purchase frequency where it would be difficult to match clicks with specific order times. Luckily, our product is a durable product (industrial door) with low order frequency per buyer for whom the matching of identified clicks with orders was easy. Additionally, there is no censoring problem in the matching given that we have the entire sales records for matching with the clickstream data.

Table 1 Summary Statistics

Variables	Mean	St. Dev.	Min	Max
Ordering Behavior				
Order	0.015	0.119	0	1
Order amount (\$)	449.26	6249.27	0	286,567.9
Order lead time (days)	89.28	103.31	0	438
General Click Measures				
Visit frequency	1.87	2.07	1	30
Average time length (seconds)	229.97	494.51	0.33	10,879.5
Average pages per visit	5.23	9.03	0.23	314.5
Detailed Click Measures				
Contact me	0.11	0.31	0	1
Contact distributor	0.05	0.21	0	1
Become distributor	0.003	0.057	0	1
Reach thanks page	0.03	0.18	0	1
Historical Ordering Behavior				
Historical order	0.04	0.19	0	1
Historical order amount (\$)	1867	17791	0	642,375
Industry Control Variables				
Chemistry industry	0.01	0.11	0	1
Food industry	0.02	0.14	0	1
Distribution industry	0.01	0.09	0	1
Manufacturing industry	0.04	0.19	0	1
Pharmaceutical industry	0.02	0.15	0	1
Transportation industry	0.01	0.10	0	1
Automobile industry	0.01	0.12	0	1

3.2. Variable Definitions

We use the (binary) indicator variable *order* as our dependent variable to denote whether the customer who clicked did purchase or not, *order amount* as a dependent variable to denote the monetary ordering amount, and *order lead time* as a dependent variable to denote the elapsed time between order placement and last time the customer visited the website.

We have four different kinds of variables that comprise our explanatory variables. First, we have “general clickstream measures” which concern data measured at a rather general level of the clickstreams. They represent the information at the level of the session which is defined as a single visit to the website. *Visit frequency* is among the most often used metrics in the e-commerce literature (cf. Moe and Fader 2004). Unlike typical e-commerce clickstream data, one characteristic of our clickstream data is that customers typically returned (if they did return) to the website after some time in the order of “days.” For the few cases of multiple sessions within a day, we aggregated these sessions within a day as one visit in our setting. *Average time length per visit* is defined as the total time stayed on the website divided by *visit frequency*, and *average number of pages per visit* is defined similarly.

Second, we have “detailed clickstream measures” that indicate whether some specific pages were visited or not. There are essentially two categories of the web pages on the firm’s website: one category of pages presents product information, and the other category shows the contact information if visitors want to contact the company, distributors, or even to become distributors.

Intuitively, we expect the pages of contact information to be more “important” than others for our information extraction purpose. Indeed, there is a lot of variation in terms of whether these contact-information pages were visited or not by the visitors. We use indicator variables to account for this variation. In particular, variables *contact me*, *contact distributor*, *become distributor*, *reach thanks page* are the ones we use to keep track of detailed clickstream information.

Third, given that new customers may derive more informational value from web browsing than existing customers, we have “historical order information” about each visitor and the dummy variable *historical order* is used to indicate whether this is an existing customer (i.e., a web visitor who has purchased before visiting the website). *Historical order amount* denotes the cumulative amount in U.S. dollars of previous orders.

Finally, some “company demographics variables,” i.e., industry control variables, are at our disposal. We include company industry type variables to control for the heterogeneity in the *latent* probability of ordering the products. The variables, *chemistry*, *food*, *distribution*, *manufacturing*, *pharmaceutical*, *transportation*, and *automobile* are used as controls for industry types. Obviously there are companies not belonging to any of these industries. It should be recognized that these control variables take into account the heterogeneity among visitors to some degree, given that all companies in the same industry are treated as homogenous. Given that our data does not allow us to pick up the customized features to individual customers, we can only treat the products as homogenous. However, the industry type controls for the heterogeneity to a certain degree. Table 1 presents the summary statistics of our data after preprocessing. From Table 1, we can indeed observe significant variations among the ordering behavior variables.

3.3. Regression Equation & Validation

We need a specific empirical prediction function $f(X_i)$ to test whether and to what extent the clickstream data is useful for demand forecasting. In the different yet related setting of e-commerce, there are a variety of prediction functions in the literature that model clicking and purchasing behavior: “conversion model” (Moe and Fader 2004), probit model (Montgomery et al. 2004), a “task-completion approach” (Sismeiro and Bucklin 2004), logit model (Van den Poel et al. 2005). We refer readers to Hui and Fader (2009) for a comprehensive literature review. The closest to ours is the seminal work by Moe and Fader (2004) who propose a conversion model and compare with several alternative models such as the logit model, duration models, Beta-Binomial, and historical conversion rates. To facilitate the comparison of the performance of the logit model versus the alternatives, we actually used their data³, and found that the logit model can perform

³ We thank them for generously sharing their data set with us.

“better” than the conversion model, even using their model evaluation criterion in their setting. To stay focused on the operational value of clickstreams, we relegate the detailed analysis to the Online Supplement. Moreover, as argued elsewhere (Van den Poel et al. 2005 for instance), the typical benefits of logit modeling are: (i) Logit modeling is well-known, simple (due to its closed-form expression) and extensively used in the literature; see, for example, Cachon et al. (2008), Draganska and Jain (2005, 2006), Train (2003) and Van den Poel et al. (2005). (ii) The ease of interpretation of logit is an important advantage over other methods. For example, the logit model can be interpreted as choices made by boundedly rational decision makers (cf. Huang et al. 2011 and references therein). For justifications and limitations of logit models, readers are referred to Cheu et al. (2009). (iii) Levin and Zahavi (1998) have shown that logit modeling provides good and robust results in general comparison studies.

We thus adopt a logit model as our prediction function f , which stems from the random utility model where we assume customer i 's outside option has normalized utility zero while purchasing yields utility:

$$U_i = \Gamma X_i + \varepsilon_i, \quad (2)$$

where $X_i = [Y_i, Z_i]$ is a vector representing customer i 's “type.” Conceptually and purely for pedagogical purposes and convenience, we can decompose the customer type to two categories:

The vector Y_i includes its *general* attributes, such as its economic characteristics, the industry it belongs to (which affects the relative usefulness of product), its size, the experiences/history of using the product and so on. In our setting, Y_i includes a set of variables to capture the customer's historical ordering behavior and a dummy variable to denote which industry it belongs to.

The vector Z_i includes the attributes of customer i 's *customized* needs; for example, a customer may need the product specialized to its business setting, and this kind of product may only be some particular firms' specialization and not others'. In our setting, Z_i is “approximated” by a set of clickstream variables defined in the previous section. To incorporate (pick up) potential nonlinear effects, we also use squares of these variables. The vector Γ denotes the coefficients of X_i and is to be estimated.

The error terms ε_i represent the un-explained variation from X_i . Under the assumption that the error terms in (2) are independently and identically distributed with the type-I extreme value distribution, the probability P_i that customer i purchases from the firm is given by the logit demand formula (McFadden 1974, 2001):

$$P_i \equiv f(X_i) = \frac{e^{U_i}}{1 + e^{U_i}}. \quad (3)$$

We now state our hypotheses to test how the clickstream data can be useful for demand forecasting. Then, we present the empirical results.

The first hypothesis is to test whether the clickstream data can be used as advance demand information:

HYPOTHESIS 1: *Visitor online behavior, as defined by the general clickstream measures and the detailed clickstream measures, is significantly correlated with offline ordering probability/propensity.*

Demand/order lead time plays an important role in operations management. While past research almost exclusively focused on predicting purchase probabilities, we also investigate whether we can use clickstream data as advance demand information to predict the *timing* of purchase. Knowing the *order lead time* (i.e., the time difference between the time of ordering and the most recent time of clicking) is beneficial for cost reduction in operations management. From a psychological perspective, a more frequent visitor would be more anxious to place orders to satisfy her need. Hence, we want to test our second hypothesis:

HYPOTHESIS 2: *Order lead time is negatively and significantly correlated with visit frequency.*

We are also interested in whether click information is useful for predicting the ordering amount as well:

HYPOTHESIS 3: *Online clicking behavior is significantly correlated with offline ordering amount.*

It is important to keep in mind that, all our hypotheses are about “correlation” rather than “causality.” Establishing the causality has been difficult in the literature, and we are not aware of any study in establishing whether clicking causes purchasing or it is vice versa. While our structural estimation results in §5.2 will help explain clicking *causes* purchasing, our data does not allow us to fully establish such a causal relationship. That requires expensive field experiments for future research, and we do not pursue that throughout the paper.

Now we are ready to present and discuss our regression results to test our hypotheses. Table 2 shows the logit regression results. From the Wald test result, our logit regression model is significant at level 0.00%. Some of the general click variables and detailed page-specific variables are statistically significant, which indicates that we shall accept HYPOTHESIS 1, i.e., visitor online click behavior is indeed providing the firm useful information to predict future ordering probabilities.

We find that *visit frequency* is positively significant at 1% level. More frequently visiting the website indeed reveals higher probability of ordering.

Table 2 also shows that the detailed click variable, *contact distributor*, is significant for predicting ordering probability. We conclude that detailed click behavior, besides general click behavior, is also useful for predicting ordering probability.

Table 2 Logistic Regression Results (Dependent Variable: Order)

Variable	All Customers	New Customers	Existing Customers
General Click Measures			
Visit frequency	0.1991** (0.0768)	0.3662** (0.1857)	0.1603* (0.0909)
Average time length	0.0015 (0.0013)	0.0038** (0.0018)	-0.0019 (0.0013)
Average pages per visit	0.0261 (0.0631)	0.1822 (0.1600)	0.0038 (0.0984)
Square of average time	-1.18e-06* (6.75e-07)	-2.50e-06** (1.12e-06)	-2.87e-07 (5.18e-07)
Square of average page	-0.0001 (0.0014)	-0.0073 (0.0061)	0.0021 (0.0027)
Square of frequency	-0.0032 (0.0031)	-0.0138 (0.0131)	-0.0016 (0.0035)
Detailed Click Measures			
Contact me or not	-0.4448 (0.5035)	-0.1858 (0.6882)	-0.4878 (0.6280)
Contact distributor	1.4183** (0.6102)	0.5995 (0.7922)	1.6455* (0.8580)
Reach thanks page	0.2143 (0.6075)	0.5252 (0.7969)	-0.1453 (0.8500)
Historical Ordering Behavior			
Industry Control Variables	Yes	Yes	Yes
Industry Control Variables			
Chemistry industry	0.9892 (0.9689)	1.7135** (0.8633)	
Food industry	-0.2017 (0.6328)	0.7696 (1.0808)	-0.4182 (0.6300)
Distribution industry	-0.8215 (1.4598)	1.4791 (0.9667)	
Manufacturing industry	0.0273 (0.5091)	0.2447 (1.0582)	-0.0332 (0.5399)
Pharmaceutical industry	-0.9083 (0.6043)		-0.7111 (0.7664)
Transportation industry	0.6167 (1.0601)		1.084 (1.6797)
Automobile industry	0.5396 (0.6325)	1.1297 (1.0428)	0.3521 (0.6514)
Constant	-6.0673*** (0.3189)	-7.3805*** (0.7182)	-1.5987*** (0.5235)
Pseudo R^2	0.3720	0.1106	0.1433

Notes. Standard errors are reported in parentheses. * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

More interestingly, from Table 3, not only does *visit frequency* convey useful information about ordering *probability*, it also provides us relevant information about the *timing* of future ordering. Indeed, if a visitor frequently visits the website, this visitor may be anxious to buy some products in near future. Hence, its order lead time may be shorter than others' *ceteris paribus*. Table 4 shows the Tobit regression results using order lead time as the non-negative dependent variable and all the other variables as explanatory variables, from which we can see *visit frequency* and *square of frequency* are significant. Hence, HYPOTHESIS 2 is accepted.

From Table 4, we can see *visit frequency*, *square of frequency*, *contact distributor*, and *historical order amount* are significantly and positively associated with *order amount*. Therefore, we can use *visit frequency* to predict both ordering probability, amount and the timing. This empirical finding is quite intriguing: It confirms that clickstream data provides advance demand information.

Table 3 Regression Results: Lead Time as the Dependent Variable

Variable	Tobit Coefficient	Variable	Tobit Coefficient
General Click Measures		Industry Control Variables	
Visit frequency	-14.7595** (5.5955)	Chemistry industry	42.6359 (79.7047)
Average time length	0.2627 (0.2559)	Food industry	57.0480 (47.0998)
Average pages per visit	-8.0866 (10.4535)	Distribution industry	41.30551 (113.3707)
Square of average time	-0.0004 (0.0003)	Manufacturing industry	-11.1666 (37.8636)
Square of average page	-8.0866 (10.4535)	Pharmaceutical industry	-31.2187 (81.0086)
Square of frequency	0.3804* (0.2084)	Transportation industry	-102.8753 (110.0317)
Detailed Click Measures		Automobile industry	-80.1476 (58.6480)
Contact me or not	33.4908 (51.6186)	Constant	155.4616*** (34.2613)
Contact distributor	-30.56358 (58.31586)		
Become a distributor	-13.7619 (78.5171)		
Reach thanks page	-15.6899 (47.6612)		

Notes. Standard errors are reported in parentheses. * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Pseudo $R^2 = 0.0207$.

Table 4 Regression Results: Order Amount as the Dependent Variable

Variable	Tobit Coefficient	Variable	Tobit Coefficient
General Click Measures		Historical Ordering Behavior	
Visit frequency	14149.12*** (2499.07)	Historical order amount	0.35*** (0.09)
Average time length	31.26 (41.87)	Industry Control Variables	
Average pages per visit	3121.84 (2416.55)	Chemistry industry	32449.68 (23724.33)
Square of average time	-0.03 (0.03)	Food industry	25410.78 (18544.97)
Square of average page	-70.55 (77.39)	Distribution industry	-15746.66 (39951.98)
Square of frequency	-328.02** (101.40)	Manufacturing industry	18659.26 (15424.23)
Detailed Click Measures		Pharmaceutical industry	-26670.31 (33357.14)
Contact me or not	-20127.95 (16309.35)	Transportation industry	17553.83 (32321.93)
Contact distributor	34150.85* (18681.09)	Automobile industry	28054.29 (22635.23)
Become a distributor	15757.68 (37431.07)	Constant	-218810.60*** (23605.47)
Reach thanks page	14755.32 (17403.75)		

Notes. Standard errors are reported in parentheses. * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Pseudo $R^2 = 0.0816$.

Table 2 also shows that the *average time length* stayed online is not significant for predicting *ordering probability*. This finding is somewhat counterintuitive. Suppose we see two visitors online, one staying very long with just a few visits, and the other visiting many times, but with short staying time each visit. Who has a higher probability of ordering *ceteris paribus*? Our results simply suggest that the second visitor is more likely to order in the future. However, as will be discussed, for the sub-population of new customers, *average time length* is significant, as shown in Table 3.

Table 2 shows the results for new customers and existing customers separately. One implication is that these two classes of customers indeed should be treated differently, in terms of linking their click behavior to their ordering probability. For new customers, *average time length* stayed online is significant to predict *ordering probability*. In addition, the relationship takes a *quadratic* form, i.e., the positive relationship trend stops at some critical point above which the relationship changes to be negatively significant. This finding confirms our intuition: Staying long online is not necessarily a good sign. For existing customers, however, there is no such significant relationship. The reason could be explained as follows: Compared with new customers, existing customers have already ordered some products before, and thus may already know enough information about the firm and the products. Hence, they probably do not need to spend much time online to collect information for purchasing decision-making. We can also see that *visit frequency* is just marginally significant (significant at level 10%) from Table 3.

4. Operational Value of Clickstream Data

In the previous section, we have provided affirmative statistical evidence that the clickstream data is useful for operational forecasting in terms of advance demand information. In this section, we will discuss what predictors from the clickstream data companies should track, and evaluate the operational value of the clickstream data based on the theoretical model in §2 and empirical analysis in §3.

Which predictors should be tracked? Although the findings here are only for a specific company, the methods are easily generalizable. In general, companies should first conduct a similar empirical study and estimate the statistical significance of both general click measures and detailed click measures as we did. This will reveal which predictors are most statistically significant for the specific setting during that specific time period. (Indeed, if seasonality is perceived to be significant, the empirical study and any parametric estimation should be performed repeatedly per season.) For example, in our setting, *visit frequency*, *average time length* and *contact distributor* are three key predictors from Table 2. This suggests that the company we have interacted with should definitely track these measures.

To illustrate how our approach and the dynamic flow equation (1) works, we now discuss how the operational forecasting process can be simulated based on our data sets. As a simple heuristic and representative example, we classify the visitors based on whether their *visit frequency* is more than 4 or not, given that those visitors who visited less than 4 times have a negligible purchasing probability on average according to our data. Hence, we effectively assume $J = 2$ classes: Visitors

who visited the website less than 4 times belong to the first class $j = 1$ having $p_t = 0$ for any t , and all the others are in class $j = 2$ having positive p_t to be estimated. We can thus omit the “class subscript” $j = 2$ in the notations for the sake of brevity. We follow two steps: (a) In each period, the new potential demand K from the new clickstreams follows a Poisson distribution with expectation μ_K which is estimated from the clickstream data. Given that the total number of visitors from the data during the one year and a half is 325, the average number of new visitors per period (i.e., three months) is approximately $\mu_K = 50$. (b) We estimate purchasing probabilities p_n and never-purchasing probabilities γ_n from the clickstream data using the empirical distribution of the click lead time: The mean purchasing probability for the visitors is 0.1046. There are $69/87 = 79.3\%$ percent of visitors whose click lead time is less than two periods based on the clickstream and sales data. Hence, all new visitors clicking in any given period will purchase with probability $p_1 = 0.1046 * 79.3\% = 0.083$ in the next period. Based on the assumption $p_t = p_1(1 - p_1 - \gamma_1)^{t-1}$, we have $p_t = \alpha p_{t-1}$ for $t = 2, 3, \dots$, where $\alpha \equiv 1 - p_1 - \gamma_1$. Hence, we can use an ordinary least squares regression (OLS) to estimate α based on the empirical distribution. We estimated that $\hat{\alpha} = 0.129$. Therefore, $p_2 = 0.129 * 0.083 = 0.011$, $p_3 = 0.129 * p_2 = 0.001$, $p_4 = 0.129 * 0.001 = 0.0002$ and $\gamma_1 = 1 - 0.083 - 0.129 = 0.788$.

For the initialization period, $t = 0$, we set $z_0 = 0$ and we have $d_1 = 0$ and $l_1 = 0$. Then $Z_1(z_0) = z_0 + K_1 - D_{j,1}(z_0) - L_1(z_0) = K_1$.

In the next period, $t = 1$, the company observes k_1 (say $k_1 = 60$) visitors on its website so that $Z_1 = z_1 = 60$. Then, $Z_2(z_1) = z_1 + K_2 - D_2(z_1) - L_2(z_1) = 60 + K_2 - D_2(z_1) - L_2(z_1)$, where $D_2(z_1)$ follows the Binomial distribution $\mathbf{B}(60; 0.083)$, and $L_2(z_1)$ follows the Binomial distribution $\mathbf{B}(60; 0.788)$. At the end of period 1, the company observes the realizations in this period, say, $k_2 = 66$, $d_2 = 5$ and $l_2 = 0$. Hence, $z_2 = k_1 + k_2 - d_2 - l_2 = 60 + 66 - 5 - 0 = 121$.

In period $t = 2$, we have the same updating: $Z_3(z_2) = z_2 + K_3 - D_3(z_2) - L_3(z_2)$, where $D_3(z_2) = D_{3,k_2} + D_{3,k_1}$. The demand D_{3,k_2} captures the conversion of the k_2 potential buyers observed in the period 1, and D_{3,k_1} comes from the k_1 potential buyers observed in period 0. It is clear that D_{3,k_2} follows distribution $\mathbf{B}(66; 0.083)$ and D_{3,k_1} follows distribution $\mathbf{B}(55; 0.083)$. Hence, $D_3(z_2)$ follows $\mathbf{B}(121; 0.083)$. Similarly, $L_3(z_2)$ follows $\mathbf{B}(121; 0.788)$. One can continue this updating for any period $t > 2$. We omit it for brevity.

Let us apply the model to the current inventory management at the company we studied. The company keeps inventory for a “patented part” (required for assembling an end product) that is supplied from Europe with a transportation lead time of three months. The company procures this component every three months, which we model as one “period” using Figure 1 in §2. The supply

Table 5 Robustness Check of the Operational Value

c	h	b	p_1	γ_1	cost reduction in percentage
80	40	400	0.05	0.85	3.65%
80	40	400	0.08	0.92	5.11%
80	40	400	0.1	0.8	5.94%
80	40	400	0.17	0.73	6.03%
80	40	400	0.2	0.7	6.12%
80	40	400	0.25	0.65	6.95%
80	40	400	0.3	0.6	7.16%
100	50	500	0.08	0.79	4.57%
110	40	500	0.08	0.79	4.29%

lead time is one period. The “demand lead time” (Hariharan and Zipkin 1995, Gallego and Özer 2001, Tan et al. 2007, Özer 2011 and references therein) is approximately zero, since customer demand is satisfied in less than two weeks. (The company can assemble-to-order within two weeks if all required components are available.)

Before quantifying the operational value in terms of cost reduction, we can first demonstrate how clickstream data improves operational forecasting by reducing demand uncertainty. We compare the variance of demand when clickstream data is not utilized versus that when clickstream data is utilized. Without clickstream data, the company can only use its prior demand distribution. Let D_L be the lead time demand without clickstream data utilized, then we have $\mathbb{E}D_L = p_1\mathbb{E}Z$ and $Var(D_L) = p_1(1 - p_1)\mathbb{E}Z + p_1^2Var(Z)$, where Z is the total number of potential buyers expressed in flow equation (1). With clickstream data utilized, however, the company can use observed clickstreams to update its demand forecast. Let D_{L1} be the lead time demand with clickstream data utilized, then $\mathbb{E}D_{L1} = \mathbb{E}D_L = p_1\mathbb{E}Z$. Invoking the law of total variance, we obtain $Var(D_{L1}) = \mathbb{E}[Var(D_{L1}(Z))|Z] + Var[\mathbb{E}(D_{L1}(Z))|Z] = p_1(1 - p_1)\mathbb{E}Z + p_1^2(1 - p_1)^2Var(Z)$. It is clear that $Var(D_{L1}) < Var(D_L)$. Using the estimated parameters from our data set, we computed $Var(D_L) \approx 4.76$ and $Var(D_{L1}) \approx 4.04$. Hence, clickstream data improves the “accuracy” of demand forecasting. However, to evaluate the operational impact of this improvement, we use the dynamic inventory control model presented in §2.

We used the following parameters: $c = 80$, $h = 0.5c$, $b = 5c$, $T = 4$ and $\beta_0 = 0.95$. We solved the dynamic programming problem based on backward induction, and we found that the annual expected cost reduction is 4.6% for these parameters. Given that these parameters are approximations, to test the robustness of the result with respect to the “accuracy” of these estimated parameters, we performed a numerical study by varying them within a reasonable neighborhood of the parameters used here. Table 5 summarizes the results as we vary the parameters in the neighborhood (to account for parameter inaccuracies), which suggest that the cost reduction is typically larger than 3%.

5. Robustness Checks

In previous sections, we used a simple logit model as our prediction function in the empirical analysis, based on which we quantified the operational value of the clickstream data. Note that there are at least two limitations: First, all visitors within each industry share the same coefficients for click variables, although we used demographic variables to take into account visitor heterogeneity. Second, the standard logit model is a *static* model which cannot capture any dynamic clicking and purchasing behavior that may potentially exist. In this section, we conduct robustness checks to partly address these limitations.

5.1. Customer Heterogeneity

In this section, we conduct a robustness check by incorporating more customer heterogeneity in the prediction function f . We allow *heterogeneity* among the coefficients of click variables even *within* each industry to check whether our previous findings are robust.

Random-coefficient logit models generalize the standard logit model by allowing coefficients to vary randomly over individuals rather than being fixed. The models do not exhibit the restrictive independence of irrelevant alternatives (IIA) property of the standard logit. As shown in McFadden and Train (2000), any pattern of substitution can be represented arbitrarily closely by a random-coefficient logit model. As discussed in Train and Revelt (1998) and Train (2003), random-coefficient logit models can take different forms in different applications, and their commonality arises in the integration of the logit formula over the distribution of unobserved random parameters. Specifically, the utility U_i for individual i can be written as $U_i = \beta_i X_i + \epsilon_i$, where β_i is a vector of coefficients that is *unobserved* for each individual i and *varies randomly* over each individual representing each individual's "tastes," and ϵ_i is an unobserved random term that is distributed i.i.d. extreme value. Suppose β_i has density $f(\beta_i|\theta^*)$ where θ^* are the (true) parameters of this distribution. Then, conditional on β_i , the probability that individual i purchases is the standard logit: $L_i(\beta_i) = \frac{e^{\beta_i X_i}}{1 + e^{\beta_i X_i}}$.

The unconditional probability is the integral of the conditional probability over all possible values of β_i : $T_i(\theta^*) = \int L_i(\beta_i) f(\beta_i|\theta^*) d\beta_i$. Maximum likelihood estimation requires the probability of each sampled individual's observed purchase. Let $I(i) \in \{0, 1\}$ indicate whether individual i purchased or not. Then the unconditional probability for the observed purchase is: $P_i(\theta^*) = \int L_{I(i)}(\beta_i) f(\beta_i|\theta^*) d\beta_i$. The log-likelihood function is $LL(\theta) = \sum_i \ln P_i(\theta)$.

Exact maximum likelihood estimation is impossible since the integral cannot be calculated analytically. Following Train (2003), we approximate the probability through simulation and maximize the *simulated log-likelihood function*. Assuming the coefficients of click variables are normally distributed, we conduct the simulated maximum likelihood estimation using KNITRO-MATLAB

Table 6 Random-coefficient Logit with Clickstream Coefficients Normally Distributed

Variable	Mean	Standard Deviation
General Click Measures		
Visit frequency	0.4281*** (0.0850)	0.0024 (0.0545)
Average time length	0.0012 (0.0020)	0.0014 (0.0017)
Average pages per visit	0.3746** (0.1863)	0.1141 (0.0982)
Square of average time	0.0000 (0.0000)	0.0000 (0.0000)
Square of average page	-0.0276* (0.0148)	0.0073* (0.0047)
Square of frequency	-0.0049 (0.0042)	0.0001 (0.0021)
Detailed Click Measures		
Contact me or not	-2.9111 (1.8364)	2.8224** (1.2267)
Contact distributor	0.6136 (0.8992)	0.0433 (1.2110)
Become a distributor	1.5293 (2.4753)	0.9147 (4.0223)
Reach thanks page	0.0543 (1.7077)	2.4760* (1.4862)
Historical Ordering Behavior	Yes	Yes
Industry Control Variables	Yes	Yes

Notes. Standard errors are reported in parentheses. * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

and report the results in Table 6. The click variables are jointly significant, suggesting that click information indeed provides useful information for predicting purchase probabilities even if visitor heterogeneity is taken care of. Furthermore, we have the same signs for these click variables as in the standard logit. This confirms the robustness of our previous results. From Table 6, we can also see that there is indeed some heterogeneity among visitors, but such heterogeneity is not significant for the majority of the click variables such as *visit frequency*.

5.2. Dynamic Customer Behavior

In this section, we ask a secondary question to the primary question answered in §3.3: Is the simple static logit model “good enough” for describing customer behavior and for converting clickstreams to advance demand information? Given the long time separation between clicks and orders, customers *could* exhibit complex, *dynamic* clicking and purchasing behavior. The data shows that customers visit the website repeatedly, presumably to learn and collect information. The information collection process may affect customer utility from purchasing the product: Only if this utility is sufficiently high do customers decide to purchase. After each visit to the website, a customer must make one of the following three choices: (1) purchase and quit visiting the website for information collection purpose⁴, (2) not purchase and quit visiting the website, (3) not purchase but return and visit the website later. A customer could be forward-looking in making her decisions in the sense

⁴We do allow customers to continue visiting the website *after* purchasing, however, we treat the purpose of these visits differently, for example, they could be driven by after-sales service enquiry etc.

that she may dynamically tradeoff the value and cost of returning: Returning to the website brings value from collecting more information, but incurs cost by postponing consumption of the product. To assess the robustness of our static logit model we want to test whether visitors are indeed forward-looking⁵ in their information collection, i.e., do they take into account *future* impacts of information collection when making *current* purchasing decisions? To answer this question, we will *extend* the simple static logit model of §3.3 to a *dynamic* model of the optimal stopping problem solved by each customer and test to what degree visitors exhibit forward-looking behavior.

After some preliminaries (relegated to Online Supplement §2.1), we can present the direct extension of the static logit model in §3.3. The state is denoted by a vector that consists of two parts: an *observed* part and an *unobserved* part by us researchers. The observed part is x , the number of visits (*visit frequency*) a customer has made. Let ε be a state variable observed by the visitor but *unobserved* by us researchers so that we have the two-dimensional state variable (x, ε) . One can treat ε as a “noise” term that includes anything that cannot be observed from the clickstream data: for example, whether the visitor visited other websites or not, the visitor’s mood when visiting, and whether this visitor has communicated with dealers or not, etc. In our setting, the company tried to contact visitors by “cold calls” offline when it started to use this clickstream tracking technology. However, this was quickly terminated for two reasons: First, most customers are large corporations and it was often impossible to track down the specific person that was visiting the website. Second, customers did not appreciate to be contacted “because we saw you visited our website.” This carried an explicit or implicit message that “big brother was watching them.” Therefore, we believe that the possible endogeneity problem from the unobserved offline interaction driven by online clicking is minimal.

We postulate that each customer solves a Markov decision stopping problem regarding purchasing or not purchasing after she clicks. We model the clicking process as a stochastic process whose transition probabilities are known by customers, but can only be observed and estimated afterwards by us researchers. Given these transition probabilities, we study the optimal purchasing decision.

The choice set for each customer is $C(x) = \{0, 1\}$ and the action vector \mathbf{a} reduces to a scalar $a \in C(x)$, where $a = 0$ means the visitor does not purchase the product and $a = 1$ means the visitor decides to purchase the product. Let β be the discount factor which denotes the *visit frequency* preference⁶. At each state x , the visitor may decide to “exit” (leave the website with or without

⁵ In a variety of different settings, this is also called “strategic” in the economics, marketing, and operations literature (e.g., Rust 1987, Nair 2007, Aviv and Pazgal 2008, to name just a few).

⁶ The discount factor β is a transformation from the conventional time preference factor β_0 , which denotes the present value of a utility unit received in the next time unit. If α_i denotes the probability transition from state (x, t) to $(x, t + i)$ in the right panel of Figure A-1, then we have that $\beta = \sum_i \alpha_i \beta_0^i \leq \beta_0$.

purchasing) or to “continue” and visit the website in the future. The latter brings the visitor to state $x + 1$ and increases learning but postpones consumption. Let $V_\theta(x, \varepsilon)$ denote the value function starting from the x th visit and the unobserved state ε , where θ is a vector of unknown parameters to be estimated that include the underlying transition probabilities, parameters of the utility function and the discount factor. Let $W_\theta(x, \varepsilon, a)$ be the value function conditional on taking action a from state (x, ε) .

Let $p(x_2, \varepsilon_2 | x_1, \varepsilon_1, a, \theta)$ be the Markov transition density from state (x_1, ε_1) to (x_2, ε_2) when a is selected and let \mathbb{E} be the expectation with respect to this probability measure. We postulate that each visitor is solving the optimal stopping problem

$$V_\theta(x, \varepsilon) = \max_{a \in \{0,1\}} [u(x, a, \theta_1) + \varepsilon(a) + \beta \mathbb{E}W_\theta(x, \varepsilon, a)],$$

which is equivalent to

$$V_\theta(x, \varepsilon) = \max\{u(x, 1, \theta_1) + \varepsilon(1) + \beta \mathbb{E}W_\theta(x, \varepsilon, 1), \bar{u}\},$$

since $u(x, 0, \theta_1) + \varepsilon(0) + \beta \mathbb{E}W_\theta(x, \varepsilon, 0) = \bar{u}$ is the outside best option, the utility from not purchasing the product. When $u(x, 1, \theta_1) + \varepsilon(1) + \beta \mathbb{E}W_\theta(x, \varepsilon, 1) > \bar{u}$, the customer will purchase the product at state (x, ε) and no future visits are needed for information collection and utility accumulation. Otherwise, the customer will choose not to purchase and may continue to visit the website to collect more information.

The transition probabilities $p(x_2, \varepsilon_2 | x_1, \varepsilon_1, a, \theta)$ over the two-dimensional states (x, ε) are difficult to estimate in general. To econometrically estimate them, we adopt the conditional independence assumption and the assumption of Type-1 extreme value distribution following Rust (1987), a seminal paper in the structural estimation literature. For brevity, we relegate these assumptions to the Online Supplement §2.2. Under these assumptions, we have the *dynamic* logit formula for the conditional purchasing probability

$$\mathbb{P}(a = 1 | x, \theta) = \frac{\exp\{u(x, 1, \theta_1) + \beta \mathbb{E}W_\theta(x, \varepsilon, 1)\}}{\exp(\bar{u}) + \exp\{u(x, 1, \theta_1) + \beta \mathbb{E}W_\theta(x, \varepsilon, 1)\}}. \quad (4)$$

Given that \bar{u} and θ_{11} are not separately identified, we normalize $\bar{u} = 0$. Notice that when $\beta = 0$, i.e., visitors are *myopic* and we recover the standard *static* logit model in §3.3. Hence, the *dynamic* logit model in this section is a direct generalization/extension of the standard logit.

The expected value function $\mathbb{E}W_\theta(x, \varepsilon, a)$ is the unique fixed point to the contraction mapping

$$\mathbb{E}W_\theta(x, \varepsilon, a) = T_\theta(\mathbb{E}W)(x, \varepsilon, a) \equiv \sum_{x'} \log \left[\sum_{a' \in \{0,1\}} \exp\{u(x', a', \theta_1) + \beta \mathbb{E}W_\theta(x', \varepsilon, a')\} \right] p(x' | x, a, \theta),$$

where the utility function is assumed to take the following form

$$u(x, a, \theta_1) = \begin{cases} \theta_{11} + \theta_{12}x^{\theta_{13}}, & \text{if } a = 1 \\ \bar{u}, & \text{if } a = 0. \end{cases}$$

Note that this utility function specification can capture the intuitive diminishing effect that the marginal utility of one more visit decreases as the number of visits increases when $\theta_{13} \in (0, 1)$. It also is sufficiently general that it includes the linear form as a special case commonly used in the literature. We also conjecture that the utility from learning through websites differs between new and existing customers.

While a transition probability matrix is typically pre-specified in the structural estimation literature, we allow a general matrix:

$$p(x_2|x_1, a, \theta) = \begin{cases} p_{x_1, x_2}, & \text{if } x_2 = x_1 \text{ and } a = 0 \\ p_{x_1, x_2}, & \text{if } x_2 = x_1 + 1 \text{ and } a = 0 \\ 1, & \text{if } x_2 = x_1 \text{ and } a = 1 \\ 0, & \text{otherwise.} \end{cases}$$

The parameters p_{x_1, x_2} are part of the parameter vector θ and are estimated from the data. This transition probability specification naturally captures that the observed state variable, *visit frequency*, can only remain the same or increase by one in each period. We use the maximum value of *visit frequency* observed in our data set to bound the state space. Hence, we choose the state space $X = \{1, 2, \dots, 20\}$ for new customers (whose maximum value of *visit frequency* in our data set is 19) and $X = \{1, 2, \dots, 31\}$ for existing customers (whose maximum value of *visit frequency* is 30).

We estimate the unknown parameter vector $\theta \equiv (\theta_{11}, \theta_{12}, \theta_{13}; \bar{u}; p_{..})$ by maximizing the likelihood function L given by

$$L(x_1, \dots, x_N, a_1, \dots, a_N | x_0, a_0, \theta) = \prod_{n=1}^N \mathbb{P}(a_n | x_n, \theta) p(x_n | x_{n-1}, a_{n-1}, \theta),$$

where N is the number of observations.

Following the constrained optimization approach proposed by Su and Judd (2008), we solve the constrained optimization problem using the advanced optimization software KNITRO:

$$\begin{aligned} & \max_{\theta, \mathbb{E}W} \prod_{n=1}^N \mathbb{P}(a_n | x_n, \theta) p(x_n | x_{n-1}, a_{n-1}, \theta) \\ & \text{subject to } \mathbb{E}W = T_{\theta}(\mathbb{E}W). \end{aligned}$$

On one extreme, one can estimate our model parameters visitor by visitor if there were enough repeated observations for each individual visitor. On the other extreme, one can treat all visitors as homogeneous and estimate our model parameters. Simple logit regression results suggest that

Table 7 Structural Estimation Results for New and Existing Customers (Number of Observations of new customers $N_1 = 8629$, and $N_2 = 1065$ of existing customers.)

Parameters	Coefficients (New Customers)	Coefficients (Existing Customers)
β (fixed)	0.95	0.95
\bar{u} (fixed)	0	0
θ_{11}	-17.73	-49.19
θ_{12}	11.54	45.69
θ_{13}	0.05	0.006

Notes. All parameters are statistically significant at 0.01 level.

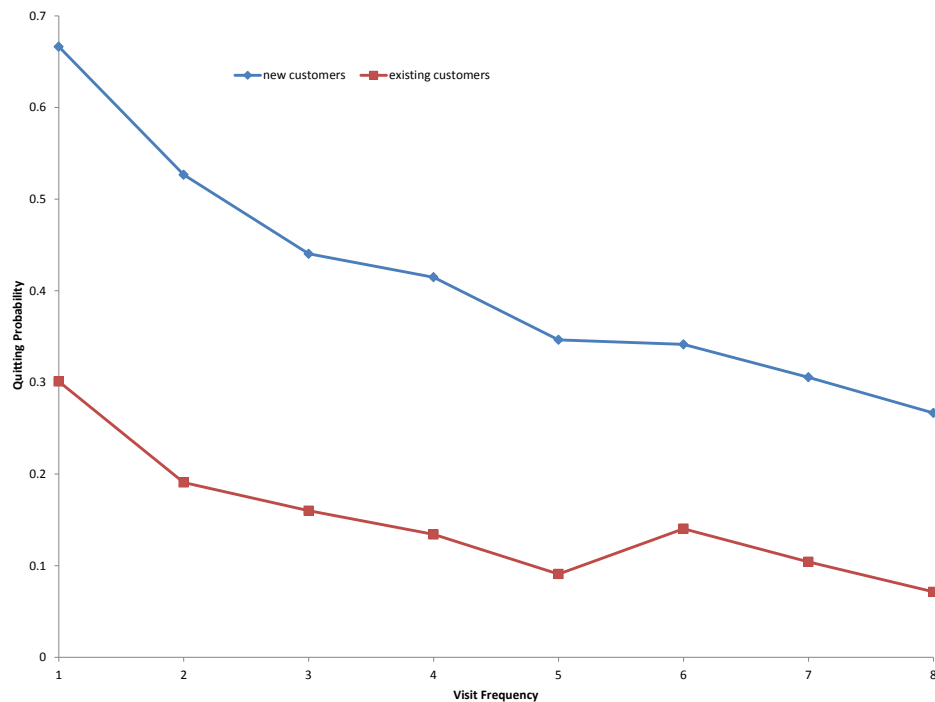
the two groups of new customers and existing customers are heterogeneous, while visitors within each group are homogenous. Hence, we choose to estimate each group separately.

Before we answer our first question regarding whether customers are forward-looking, we present some estimation examples to demonstrate that our model yields intuitive results. For the group of new customers, as a representative example, suppose we fix $\beta = 0.95$. (Estimation of the discount factor is difficult, see for example, Nair 2007, and we do not pursue it. Rather, we will perform the structural estimation for a discrete set of β in $[0, 1]$.) Then, the estimated utility function is $u(x, 1, \theta_1) = -17.73 + 11.54x^{0.05}$, and all the estimated parameters are statistically significant. The results are intuitive: (1) Information collection behavior on the website and customer utility are statistically significantly correlated, given that $\theta_{12} > 0$; (2) The postulated diminishing effect indeed exists here given that $\theta_{13} < 1$. For the group of existing customers, the estimated utility function is $u(x, 1, \theta_1) = -49.19 + 45.69x^{0.006}$, and all the parameters are statistically significant. We present these results in Table 7.

While establishing the causal relationship between clicking and purchasing has been difficult in the literature (indeed, we are not aware of any study that establishes the causal relationship) and in our setting, we can offer some insights towards this direction by comparing the estimation results for new and existing customers in Table 7. Intuitively, if clicking increases customer utility from purchasing, new customers should benefit more than existing customers given that existing customers are more familiar with the company and its products. This is indeed the case by comparing the marginal benefit from one more visit: For new customers, it is $\frac{du}{dx}|_{new} = \frac{0.577}{x^{0.95}}$, while it is $\frac{du}{dx}|_{existing} = \frac{0.274}{x^{0.99}}$ for existing customers. Hence, $\frac{\frac{du}{dx}|_{new}}{\frac{du}{dx}|_{existing}} \approx 2x^{0.04} > 2$ for all $x \geq 1$. This suggests that the benefit from one more visit for new customers is **more than double** the benefit for existing customers.

Estimation results for transition probabilities $p(x|x, a = 0, \theta)$ are plotted in Figure 2, for both new and existing customers. Interestingly, the estimated transition matrix suggests that it is more likely that visitors quit browsing without purchasing when visit frequency is small. When the visit frequency becomes larger, such quitting without purchasing probabilities become smaller. This

Figure 2 Quitting Probability versus Visit Frequency: Existing Customers Are More “Loyal”



may be treated as a “lock-in effect” of visiting the website: The more often visitors have visited the website, the less beneficial it is for them to quit without purchasing. The probability $p_{i,i}$ can be interpreted as the “quitting probability” without purchasing, then Figure 2 shows the lock-in effect. One should recognize that such interesting result is obtained endogenously, i.e., estimated from the data without imposing any restrictions on the transition probability matrix.

The quitting probabilities for new customers are systematically higher than those for existing customers. This suggests that existing customers are more “loyal” to the company since they are more likely to come back to the website in the future. In contrast, new customers appear more frivolous. It is comforting that the structural estimation confirms intuition: Existing customers have purchased from the company before and are more serious in their clicking decisions than new customers.

After affirming that our model yields intuitive results, we are ready to answer: Are visitors *forward-looking* ($\beta \approx 1$) or *myopic* ($\beta \approx 0$) in their information collection? It would be ideal to answer this question using all the statistically significant clickstream variables among X_i . However, due to computational challenges as we shall discuss, we conduct the following analysis.

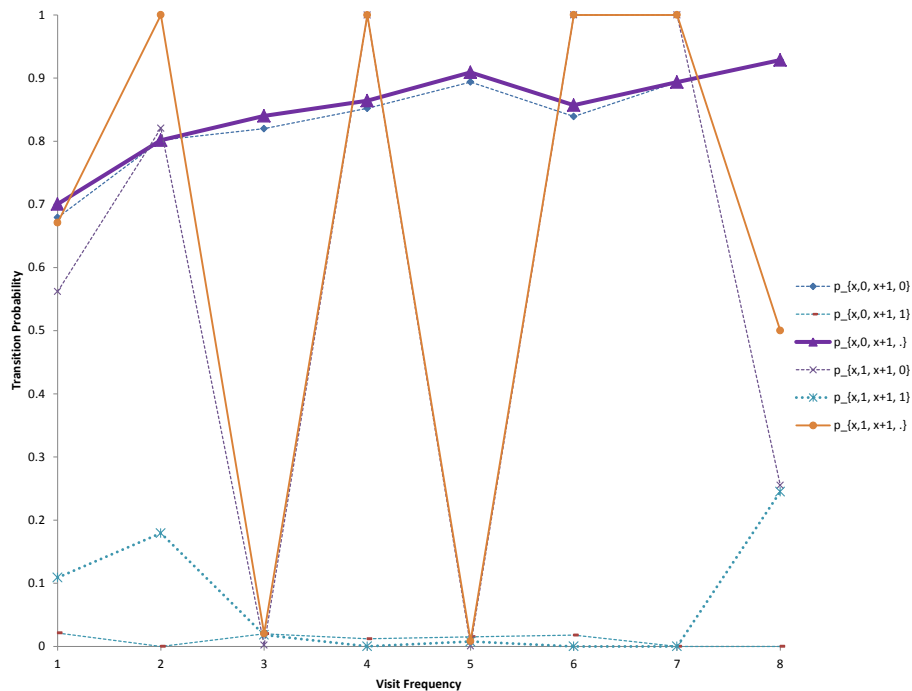
We first use a single observed state variable x , and find that the dynamic nature of customer tradeoff in purchasing we postulated is not statistically significant for both of the subgroups:

new and existing customers (see Online Supplement §2.3 for details). Using one single observable state variable x in our structural estimation model has limitations. Table 2 suggests that for new customers, *Average time length* is also statistically significant besides *visit frequency*. It would be desirable if we could include *Average time length* variable in our structural estimation model. However, this is challenging given that *Average time length* is a continuous variable; estimating its transition probability matrix would require discretization. The associated state space explosion makes the estimation prohibitive. Yet Table 3 shows that *Contact distributor* is also statistically significant for existing customers besides *visit frequency*. (For new customers, *Contact distributor* is not statistically significant in Table 2.) Given that *Contact distributor* is a binary variable, the state space only doubles and we were able to estimate the extended model.

For the group of existing customers, we use (x, δ) as the observed state vector where $\delta = 1$ if *Contact distributor*=1 in the x -th visit and $\delta = 0$ otherwise. We denote the transition probability from state (x_1, δ_1) to state (x_2, δ_2) by $p_{x_1, \delta_1, x_2, \delta_2}$. This transition matrix is a sparse matrix and it is sufficient to estimate the “return probability” $p_{x, \delta_1, x+1, \delta_2}$. We follow the same estimation procedures as before. Similar as before, we postulate that the utility function takes this intuitive form: $u(x, \delta, a = 1, \theta_1) = \theta_{11} + \theta_{12}x^{\theta_{13}} + \theta_{13}\delta$.

First, it is instructive to discuss some estimation examples. For $\beta = 1$, the estimated utility function of existing customers is $u(x, \delta, a = 1, \theta_1) = -260.94 + 257.10x^{0.001} + 2.53\delta$, and all the parameters are statistically significant. The estimated transition probabilities are shown in Figure 3. The return probability as a function of the *visit frequency* is complex. However, it is interesting that the return probability $p_{x, 0, x+1, \cdot} \equiv p_{x, 0, x+1, 0} + p_{x, 0, x+1, 1}$ (in bold) appears to be roughly increasing in *visit frequency* in the range where we have many observations, suggesting the lock-in effect discussed before persists.

Second, are the existing customers forward-looking? We conduct a likelihood ratio test as before. We obtain $LL(0) = -651.136 < LL(1) = -645.707$. The P-value=0.001 (the probability under chi-square distribution with 1 degree of freedom: $\text{Prob} > \text{chi2tail}(1, 2 * (-645.707 + 651.136))$) suggesting that the forward-looking behavior is actually statistically significant for existing customers. **This demonstrates that existing customers *do* exhibit forward-looking behavior after we incorporate “rich ingredients” such as the detailed clicktream path per visit.** This finding justifies that our postulated customer dynamic clicking and purchasing behavior is not “artificial,” indeed exists, and is worth future research investigations. New customers, however, remain myopic even in the richer model given that *Contact distributor* is not statistically significant. In our

Figure 3 Return Probability versus Visit Frequency: Lock-in Effect Persists


setting, new customers constitute the vast majority ($4982/5185 = 96\%$) of the visitors, and they can be simply treated as myopic. Hence, we conclude: while simply adopting the static logit model in §3.3 in our setting is reasonable, the fact that existing customers *do* exhibit forward-looking behavior justifies attempting the dynamic logit model here and for future research. Extensions to include *Average time length* in the structural model are currently computationally prohibitive and are “future aspirations.”

6. Discussions and Limitations

Our primary goal of this study is to show how and to what extent we can use clickstream data from non-transactional websites to improve operational forecasting and inventory management to reduce supply-demand mismatches. We first introduced a dynamic decision support model that includes clickstreams as state variables in inventory management. Second, we conducted an empirical study to identify which clickstream variables are statistically significant for demand forecasting and to estimate the extent to which including these clickstreams reduces operational costs.

Our study is motivated by practice and is aimed to guide better practice of clickstream tracking in operations management. Our model provides a practical framework to dynamically convert clickstream data into useful advance demand information for inventory management. In practice, firms should develop decision support systems using clickstream data by taking advantage of various

statistical and computer science tools, such as data mining and artificial intelligence, to enhance the prediction from the regression equation (e.g., using more sophisticated prediction function $f(X_i)$) and better extract advance demand information from the clickstream data.

Our findings must be interpreted cautiously given the limitations of our study: First, we only used the visitors who are identifiable in our clickstream data set, which can create biases for our empirical study. Companies should consider mechanisms to improve customer identification of clickstreams (e.g., use cookies, let customers sign in and provide more information, etc.). Second, considering the heterogeneity of visitors, our control variables are limited. For example, price is negotiated offline and such information is unobserved by us. While this is the best our data allows, we can take comfort given that the random-coefficient logit model further takes care of the heterogeneity to some degree. Third, we do not conduct time series analysis due to our limited observations within a short period of time. Availability of large-scale data sets for a long period of time would allow us to investigate the dynamics over time. Finally, although our models and methods can be easily generalized and applied to other settings of offline sales with informational websites, all the findings herein are based on the data from a particular industrial firm with a fixed period of visiting customers. We hope our study stimulates more research in this important, practice-driven and data-driven area.

References

- Aviv, Y., A. Pazgal. 2008. Optimal pricing of seasonal products in the presence of forward-looking consumers. *Manufacturing Service Oper. Management*. 10(3) 339-359.
- Cachon, G., C. Terwiesch, Y. Xu. 2008. On the effects of consumer search and firm entry on multiproduct competition. *Marketing Sci.* 27(3) 461-473.
- Caldentey, R., G. Vulcano. 2007. Online auction and list price revenue management. *Management Sci.* 53(5) 795-813.
- Chen, F. 2001. Market segmentation, advanced demand information, and supply chain performance. *Manufacturing Service Oper. Management* 3(1) 53-67.
- Cheu, R. L., H. Nguyen, T. Magoc, V. Kreinovich. 2009. Logit discrete choice model: A new distribution-free justification. *Soft Computing - A Fusion of Foundations, Methodologies and Applications* 13(2) 1432-7643.
- Draganska, M., D. Jain. 2005. Product-line length as a competitive tool. *J. Econom. Management Strategy* 14(1) 1-28.
- Draganska, M., D. Jain. 2006. Consumer preferences and productline pricing strategies: An empirical analysis. *Marketing Sci.* 25(2) 164-147.

-
- Fay, S., D. Mitra, Q. Wang. 2009. Ask or infer? Strategic implications of alternative learning approaches in customization. *Intern. J. of Research in Marketing* 26 136-152.
- Gallego, G., Ö. Özer. 2001. Integrating replenishment decisions with advance demand information. *Management Sci.* 47(10) 1344-1360.
- Gallego, G., Ö. Özer. 2003. Optimal replenishment policies for multiechelon inventory problems under advance demand information. *Manufacturing Service Oper. Management* 5(2) 157-175.
- Gayon, J. P., S. Benjaafar, F. de Vericourt. 2009. Using imperfect demand information in production-inventory systems with multiple demand classes. *Manufacturing Service Oper. Management* 11 128-143.
- Hariharan, R., P. Zipkin. 1995. Customer-order information, leadtimes, and inventories. *Management Sci.* 41(10). 1599-1607.
- Huang, T., G. Allon, A. Bassamboo. 2011. Bounded rationality in service systems. *Working paper*, Kellogg School of Management, Northwestern University.
- Huberman, B.A., P. T. Pirolli, J. E. Pitkow, R. M. Lukose. 1998. Strong regularities in World Wide Web surfing. *Science* 280 95-7.
- Hui, K., P. S. Fader, E. T. Bradlow. 2009. Path data in marketing: An integrative framework and prospectus for model building, *Marketing Sci.* 28(2) 320-335.
- Johnson, E. J., S. Bellman, G. L. Lohse. 2003. What makes a web site “sticky”? Cognitive lock in and the power law of practice. *Journal of Marketing* 67 62-75.
- Lariviere, M., J. A. Van Mieghem. 2004. Strategically seeking service: How competition can generate Poisson arrivals. *Manufacturing Service Oper. Management* 6(1) 23-40.
- Levin, N., J. Zahavi. 1998. Continuous predictive modeling: A comparative analysis. *Journal of Interactive Marketing* 12(2) 5-22.
- McFadden, D., K. Train. 2000. Mixed MNL Models for discrete response. *Journal of Applied Econometrics* 15 447-470.
- McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In Zarembka P. (ed) *Frontiers in econometrics*. Academic Press, New York, 105-142.
- McFadden, D.. 2001. Economic choices. *American Economic Review* 91 351-378.
- Moe, W., P. S. Fader. 2004. Dynamic conversion behavior at e-commerce sites. *Management Sci.* 3 326-335.
- Montgomery, A. L. 2001. Applying quantitative marketing techniques to the internet. *Interfaces* 31(2) 90-108.
- Montgomery, A. L., K. Srinivasan. 2003. Learning about customers without asking, Nirmal Pal and Arvind Rangawamy (eds.), *The Power of One-Leverage Value from Personalization Technologies*, eBRC Press, Penn State University.
- Montgomery, A., S. Li, K. Srinivasan, J. C Liechty. 2004. Modeling online browsing and path analysis using clickstream data. *Marketing Sci.* 23(4) 579-585.

- Nair, H. 2007. Intertemporal price discrimination with forward-looking consumers: Application to the US market for console video-games. *Quantitative Marketing and Economics* 5(3) 239-292.
- Özer, Ö. 2011. Inventory management: Information, coordination and rationality. Chapter 13 in *Handbook of Production Planning*, Ed. K. Kempf, P. Keskinocak and R. Uzsoy, 321-365.
- Özer, Ö., W. Wei. 2004. Inventory control with limited capacity and advance demand information. *Operations Research* 52 988-1000.
- Porteus, E. 1990. Stochastic inventory theory. Chapter 12 in *Handbooks in Operations Research and Management Science* 2 605-652.
- Raman, A., M. Fisher. 1996. Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research* 44(4) 87-99.
- Rust, J.. 1987. Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica* 55(5) 999-1033.
- Sheskin, D.. 2004. *Handbook of Parametric and Nonparametric Statistical Procedures*. Third Edition, Chapman & Hall/CRC.
- Sismeiro, C., R. E. Bucklin. 2004. Modeling purchase behavior at an e-commerce web site: A task completion approach. *Journal of Marketing Research* 306-323.
- Su, C. L., K. L. Judd. 2008. Constrained optimization approaches to estimation of structural models. Working paper, University of Chicago.
- Tan, T., R. Gullu, N. Erkip. 2007. Modeling imperfect advance demand information and analysis of optimal inventory policies. *Eur. J. Oper. Res.* 111 897-923.
- Train, K. 2003. *Discrete Choice Methods with Simulation*, Cambridge University Press, New York.
- Train, K., D. Revelt. 1998. Mixed logit with repeated choices: Households' choices of appliance efficiency level. *Review of Economics and Statistics* 4 647-657.
- Van den Poel, D., W. Buckinx. 2005. Predicting online-purchasing behavior. *Eur. J. Oper. Res.* 166 557-575.
- Wang, T., B. Toktay. 2008. Inventory management with advance demand information and flexible delivery. *Management Sci.* 54(4) 716-732.

Online Supplement

for “Clickstream Data and Inventory Management: Model and Empirical Analysis”

This Online Supplement comprises of two parts: The first part provides partial support and motivation for adopting the logit model in §3.3. The second part contains supporting materials of the dynamic logit model presented in §5.2.

Online Supplement 1. Logit versus “Conversion Model”

One of our motivations for exploring a visitor economic/behavioral model stems from studying the “conversion model” using clickstream data in Moe and Fader (2004) (from now on MF), a seminal paper of e-commerce clickstreams. Although the setting there is different from ours, we argue that our approach, i.e., the random-utility model, can perform “better” than the conversion model. This result should be treated as a *byproduct* of our study (which focuses on demonstrating and quantifying the operations value of the clickstream data in an offline setting). For clarity, we first briefly summarize their work, and then demonstrate how a logit model can perform better.

MF adopt the novel idea from Schmittlein and Morrison (2003), who use a “fraction” to model the success probability of in vitro fertilization (IVF). Specifically, Schmittlein and Morrison (2003) use x to represent the impact of success factors and y to represent the factors making the IVF cycle likely to fail. The probability of success of the IVF trial is $p = \frac{x}{x+y}$. Following Schmittlein and Morrison (2003), MF let p_{ij} be the probability of visitor i purchasing at visit j , then $p_{ij} = \frac{V_{ij}}{V_{ij} + \tau_{ij}}$, where V_{ij} denotes the net effect of visits since last purchase, and τ_{ij} denotes the purchasing threshold. MF assume that the net visit effect, V_{ij} consists of two components: a baseline propensity to buy, v_{i0} , that applies at every visit, and incremental effects, m_{ij} that accumulate over all visits that have occurred since that last purchase. In a customer’s first purchasing cycle, they have $V_{ij} = v_{i0} + m_{i1} + m_{i2} + \dots + m_{ij}$ for household i who has made j non-purchase visits. To accommodate different forms of heterogeneity, they assume that the baseline purchasing propensity v_{i0} is gamma distributed across the customer population with shape parameter r_v and scale parameter γ . They also assume that the visit impacts and the purchasing thresholds vary across customers in accordance with a gamma distribution, such that $m_{ij} \sim \text{gamma}(\mu, \gamma)$ and $\tau_{ij} \sim \text{gamma}(r_\tau, \gamma)$. To take the evolving visiting effects into account, they assume $V_{ij} \sim \text{gamma}(r_v, \gamma) + \text{gamma}(\mu_0 k^1, \gamma) + \text{gamma}(\mu_0 k^2, \gamma) + \dots + \text{gamma}(\mu_0 k^j, \gamma)$, where the parameter k ranges from zero to infinity and characterizes how visit impacts involve as customer familiarity increases.

To incorporate the evolving purchasing thresholds, they assume $\tau_{ij} \sim \text{gamma}(r_\tau \exp\{\psi x_{ij}\}, \gamma)$, where r_τ captures the initial purchasing threshold, ψ is a parameter that governs the magnitude

and direction of the dynamic process, and x_{ij} is the number of purchases that customer i has made, up to (but not including) visit j . Then, they derive the likelihood function, apply to the observed panel data, and find that their model fits the data better than existing simple models, such as the simple logistic regression. As they argued,

... model the probability of purchasing in each session as a function of: (1) the number of past visits (2) the number of past purchases (3) the number of visits since the last purchase (4) time elapsed (in days) since the last visit (5) time elapsed (in days) since the last purchase. The fit of the model (LL=-4367.79 and BIC=8791.55) is vastly inferior to that of the conversion model...

Table A-1 Replicated Logistic Regression in MF

Variable	Coef.	Std. Error	z	$P > z $
Number of past purchases	0.3775*	0.0260	14.50	0.000
Number of past visits	-0.0702*	0.0096	-7.31	0.000
Number of visits since last purchase	0.0160	0.0269	0.59	0.553
Time elapsed since last purchase	0.0065*	0.0011	5.67	0.000
Time elapsed since last visit	0.0002	0.0009	0.23	0.821
Constant	-1.9446*	0.0356	-54.55	0.000

*Statistically significant at 5% level; LL=-4367.6239; Number of parameters: $k=6$

Table A-2 "Refined" Logistic Regression

Variable	Coef.	Std. Error	z	$P > z $
Number of past purchases	0.3719*	0.0242	15.36	0.000
Number of past visits	-0.0668*	0.0076	-8.85	0.000
Time elapsed since last purchase	0.0069*	0.0009	7.53	0.000
Constant	-1.9445*	0.0329	-59.06	0.000

*Statistically significant at 5% level; LL=-4367.8022; Number of parameters: $k=4$

Table A-3 Logistic Regression Example (1)

Variable	Coef.	Std. Error	z	$P > z $
Number of past purchases	0.5789*	0.1071	5.41	0.000
Number of past visits	-0.0828*	0.0081	-10.28	0.000
$\exp(0.7 \cdot \text{Number of past purchases})$	-2.0211*	0.5957	-3.39	0.001
$0.1 \cdot \text{Number of past purchases}$	-0.8194*	0.1155	-7.09	0.000
$0.1 \cdot \text{Number of past visits}$	-0.3492*	0.0752	-4.64	0.000
Constant	0.9279*	0.5351	1.73	0.083

*Statistically significant at 5% level; LL=-4264.1101; Number of parameters: $k=6$.

Table A-4 Logistic Regression Example (2)

Variable	Coef.	Std. Error	z	$P > z $
Number of past purchases	-0.3168*	0.1470	-2.16	0.031
Number of past purchases ^{$\frac{1}{3}$}	-1.6752	1.0314	-1.62	0.104
Number of past visits ^{$\frac{1}{3}$}	1.5637*	0.1732	9.03	0.000
Number of past purchases ^{$\frac{1}{2}$}	3.0963*	1.1321	2.73	0.006
Number of past visits ^{$\frac{1}{2}$}	-1.2586*	0.1200	-10.48	0.000
Constant	-2.2642*	0.0513	-44.10	0.000

*Statistically significant at 5% level; LL=-4263.4862; Number of parameters: $k=6$.

Table A-5 Logistic Regression Example (3)

Variable	Coef.	Std. Error	z	$P > z $
Number of past purchases ^{$\frac{1}{3}$}	17.6072*	8.3302	2.11	0.035
Number of past visits ^{$\frac{1}{3}$}	1.5613*	0.1732	9.02	0.000
Number of past purchases ^{$\frac{1}{2}$}	-4.0176	2.3120	-1.74	0.082
Number of past visits ^{$\frac{1}{2}$}	-1.2566*	0.1200	-10.47	0.000
Number of past purchases ^{$\frac{1}{4}$}	-12.4882*	6.0602	-2.06	0.039
Constant	-2.2642*	0.0513	-44.10	0.000

*Statistically significant at 5% level; LL=-4263.7388; Number of parameters: $k=6$.

Table A-6 Logistic Regression Example (4)

Variable	Coef.	Std. Error	z	$P > z $
Number of past purchases	-0.3092*	0.1318	-2.35	0.019
Number of past purchases ^{$\frac{1}{2}$}	2.5726*	0.7608	3.38	0.001
Number of past visits ^{$\frac{1}{2}$}	-0.8748*	0.0796	-10.99	0.000
Number of past purchases ^{$\frac{1}{4}$}	-1.1504	0.6741	-1.71	0.088
Number of past visits ^{$\frac{1}{4}$}	1.1986	0.1322	9.06	0.000
Constant	-2.2663*	0.0516	-43.95	0.000

*Statistically significant at 5% level; LL=-4263.8264; Number of parameters: $k=6$.

Table A-7 Logistic Regression Example (5)

Variable	Coef.	Std. Error	z	$P > z $
Number of past purchases ^{$\frac{1}{4}$}	-12.4882*	6.0602	-2.06	0.039
Number of past purchases ^{$\frac{1}{2}$}	-4.0176	2.3120	-1.74	0.082
Number of past purchases ^{$\frac{1}{3}$}	17.6072*	8.3302	2.11	0.035
Number of past visits ^{$\frac{1}{2}$}	-1.2566*	0.1200	-10.47	0.000
Number of past visits ^{$\frac{1}{3}$}	1.5613*	0.1732	9.02	0.000
Constant	-2.2642*	0.0513	-44.10	0.000

*Statistically significant at 5% level; LL=-4263.7388; Number of parameters: $k=6$.

For the conversion model, LL=-4264.25, BIC=8578.81. We replicated the logit regression and found that the *number of visits since last purchase* and the *time elapsed since last visit* are not

Table A-8 Logistic Regression Example (6)

Variable	Coef.	Std. Error	z	$P > z $
Number of past purchases	0.3942*	0.0529	7.45	0.000
Number of past visits	-0.0622*	0.0160	-3.89	0.000
Number of past purchases ^{$\frac{1}{3}$}	0.9066*	0.1058	8.57	0.000
Number of past visits ^{$\frac{1}{3}$}	-0.2726	0.1817	-1.50	0.134
exp(0.7*Number of past purchases)	-1.6202*	0.1968	-8.23	0.000
0.1Number of past visits	-0.6419*	0.2037	-3.15	0.002

*Statistically significant at 5% level; LL=-4263.2937; Number of parameters: $k=6$.

significant, as shown in Table A-1. Eliminating the two variables improves the BIC without affecting LL very much, as shown in Table A-2.

Note that we have limited information from the data set, given that no detailed click information such as which web pages are visited, etc., is provided. Hence, we question the appropriateness of simply aggregating or summing the past visits and purchases as “attributes” (explanatory variables). Intuitively, the visits and purchases may have a diminishing effect on purchasing as the numbers increase, and we can re-scale or discount the numbers to incorporate such effect to some degree. For example, we can use various functions of the number of visits or purchases to approximate the “attributes.” We can easily pick up some discounting functions in our regression. Tables A-3-A-8 show some simple and readily found examples that are no worse than the conversion model in terms of the criterion, LL and BIC, used in MF. Our general two-stage approach would be: (1) Exploring: try different common nonlinear forms of raw data and do maximum likelihood estimation using the logit type model; (2) Exploiting: fix some parameters (to reduce the number of parameters) from the first-stage result, then do logit (or maximum likelihood estimation in general) using the explanatory variables from the first stage. This method can be better than, or at least no worse than the conversion model proposed in MF using their criterion. One of many advantages of using logit-type models is that it is a behavioral model coming from rational utility maximization.

As MF argued and we agree, the conversion MF model has some compelling features such as allowing the heterogeneity across visitors, heterogenous effects of visits and purchases, and non-stationarity of these effects. However, this model does not explicitly capture each visitor’s underlying purchasing decision-making behavior. We know the probability of purchasing for human beings should be fundamentally different from the probability of IVF’s success which is mainly determined exogenously out of human-beings’ control. As we have shown, the simple logit model can perform as well as the conversion model in terms of LL and BIC if we use carefully-chosen explanatory variables,

and the advantage for the logit model is that it is based on the commonly used random utility model, a behavioral model based on utility maximization.

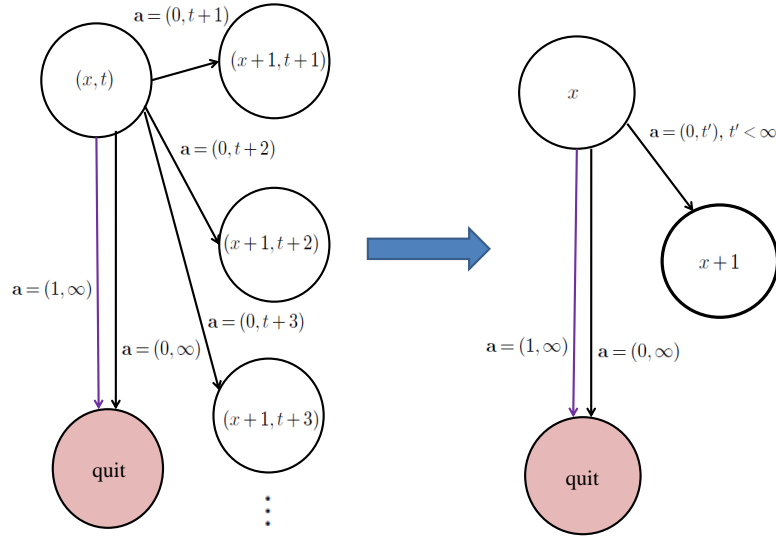
Online Supplement 2. Supporting Materials for Dynamic Logit

Online Supplement 2.1. Preliminaries

To intuitively understand how the extension of the simple static logit model works, we start with a discrete-time framework where each customer makes her clicking and purchasing decisions at each “state” observable to us researchers. The state here is a vector (x, t) , where x is the number of visits (i.e., *visit frequency*) a customer has made up to period t . At state (x, t) , a customer must make two decisions: clicking and purchasing. Formally, let the action be $\mathbf{a} = (a_1, a_2)$, where a_1 denotes the purchasing decision and a_2 denotes the clicking decision. $a_1 = 1$ means she purchases the product and $a_1 = 0$ means she does not. a_2 denotes the time at which she will visit the website again in the future. Clearly $a_2 = \infty$ means she will never come back, i.e., quit visiting the website. We depict the transition diagram in the left panel of Figure A-1 in the Online Supplement. Typically, the intended future visit time a_2 is subject to some exogenous randomness: other factors may crop up that change (often delay) the actual visit time. Analyzing and estimating this model with highly connected two-dimensional state space is challenging, so we transform this model to a simplified version by dropping the time variable in the state description as depicted in the right panel of Figure A-1. Effectively, we “aggregate” all the states $(x + 1, t)$ for $t < \infty$ to a single state $x + 1$.

In the diagram depicted in the right panel of Figure A-1, there remain two decisions for each customer: clicking and purchasing. A customer fully controls her purchasing decisions based on her own utility maximization. But whether she actually will visit in the future is subject to uncertainty: Unanticipated events could make the transition to the next state $x + 1$ *probabilistic*. Hence, clicking decisions result from both endogenous rational utility-maximization and exogenous factors. Another interpretation is that, even if customers were perfectly rational in their clicking decisions, we researchers (as outside observers) can only treat their decisions as probabilistic choices because of unobserved noises and heterogeneity, and estimate them. This treatment is consistent with the literature, and we refer readers to Huberman et al. (1998), a seminal paper on modeling and estimating web browsing behavior.

Let $p_{x,x+1}$ be the probability that a visitor will return to the website given that she already visited x times and that she has not purchased the product yet. To isolate and focus on customer purchasing behavior, we will separate the clicking decisions from the purchasing decisions. We will estimate the clicking probabilities $p_{x,x+1}$ from the data. Given these clicking probabilities, we will then study how a visitor at each state makes her optimal purchasing decision. This approach simplifies the

Figure A-1 Transformation of Transition Diagrams

dynamic programming formulation to a stopping problem in a Markov chain. Using state-of-the-art optimization software, we can simultaneously estimate the clicking probabilities and the structure of purchasing decision. In addition to making the structural estimation feasible, this approach of separating clicking and purchasing decisions also follows the literature (see, e.g., Huberman et al. 1998, Montgomery 2004 and references therein, where clicking decisions are modeled independent of purchasing decisions).

Online Supplement 2.2. Assumptions for Dynamic Logit

ASSUMPTION 1. *The transition density of the controlled process $\{x, \varepsilon\}$ factors as*

$$p(x_2, \varepsilon_2 | x_1, \varepsilon_1, a, \theta) = q(\varepsilon_2 | x_2, \theta_2) p(x_2 | x_1, a, \theta_3).$$

This assumption requires two restrictions. First, the cumulative number of visits (*visit frequency*) x_2 is a sufficient statistic for the unobserved noise term ε_2 . Second, the probability density of the next *visit frequency* x_2 depends only on the current *visit frequency* x_1 and not the unobserved term ε_1 . For benefits and limitations of this assumption, see Rust (1987).

Following Rust (1987), which is used extensively in the literature such as Nair (2007), we assume $q(\varepsilon | x, \theta_2)$ is given by an multivariate Type-1 extreme value distribution

$$q(\varepsilon | x, \theta_2) = \prod_{a \in \{0,1\}} \exp\{-\varepsilon(a) + \theta_2\} \exp\{-\exp\{-\varepsilon(a) + \theta_2\}\},$$

where $\theta_2 = 0.577216$.

Online Supplement 2.3. Structural Estimation with State (x, ε)

For each customer group, we computed the log likelihood $LL(\beta)$ as a function of the discount factor at 0 and 1, and conducted the likelihood ratio test (cf. Rust 1987). For new customers, $LL(0) = -5784.4482$ and $LL(1) = -5784.4484 < LL(0)$. Hence, it is more likely that new customers are myopic. In contrast, existing customers are more likely to be forward-looking because $LL(0) = -592.457$ and $LL(1) = -592.382 > LL(0)$. Comparing the forward-looking model and myopic model in terms of the maximum likelihood values, however, we find that the P-value=0.699 (the probability under chi-square distribution with 1 degree of freedom: $\text{Prob} > \text{chi2tail}(1, 2 * (-592.382 + 592.457))$). Hence, for existing customers, the forward-looking model with positive β is not statistically significantly different from the myopic model with $\beta = 0$. This suggests that, based on our structurally estimated model, we do not have sufficient evidence to reject the hypothesis that existing customers are myopic: they do not take the expectation of future visit effects into consideration when they make purchasing decisions now. This finding suggests that the dynamic nature of customer tradeoff in purchasing we postulated is not statistically significant.