POMS PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT POMS

Vol. 23, No. 3, March 2014, pp. 333–347
ISSN 1059-1478 | EISSN 1937-5956 | 14 | 2303 | 0333
DOI 10.1111/poms.12046

Clickstream Data and Inventory Management: Model and Empirical Analysis

Tingliang Huang
Department of Management Science and Innovation, University College London, London, WC1E 6BT, UK, t.huang@ucl.ac.uk

Jan A. Van Mieghem
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208, USA, 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 can reduce the inventory holding and backordering cost by 3% to 5% in our data set.

Key words: click tracking; advance demand information; inventory theory and control; empirical research; dynamic programming; econometric analysis; big data
History: Received: January 2012; Accepted: December 2012, by Panos Kouvelis, after 2 revisions.

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 study) 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 business-to-business (B2B) settings as well as some business-to-consumer (B2C) settings fall in this category. Specifically, this study stems from our interaction with a US 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. 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 benefit 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. In particular, 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 statistically 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 several reasons. 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. Understanding consumer online browsing behavior and its value helps firms make investment decisions regarding the adoption of clickstream tracking technology. Manyika et al. (2011) report that "big data—large pools of data that can be captured, communicated, aggregated, stored, and analyzed—is now part of every sector and function of the global economy." Clickstream tracking has allowed individuals around the world to contribute to the amount of big data available to companies. Our study examines the potential operational value that clickstream data, an important type of big data, can create for companies and seeks to illustrate and quantify that value. In a concrete setting of the company, we show that using the information extracted from the clickstream data can reduce the inventory holding and backordering cost by 3% to 5% in many representative parameter scenarios. The model and empirical methods we use in our study may be useful for other companies that aim to exploit big data to gain competitive advantage.

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. While it has been confirmed in the literature that online click behavior is 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. It is reported that e-commerce sales only account for 1.2% of all retail sales.[1] Hence, the vast majority of commerce still is executed offline, and thus our research setting addresses a larger part of the economy beyond e-commerce.

Due to the procedural separation, non-transactional websites provide the opportunity for firms to react. Clearly, in an e-commerce setting like Amazon, the time lag between clicks and orders could be on the order of minutes, too short to adjust operational plans. The longer time separation between clicks and orders has an important benefit: if it exceeds the production or procurement lead time, the firm can respond to changes in advance demand information. 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 studies 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."

Our work is also related to recent empirical study in the information systems literature of using keyword search and social mentions to predict future events, based on the idea that what people are searching for today is predictive of what they will do in the future (cf. Asur and Huberman 2010, Goel et al. 2010, Joo et al. 2011, and reference therein). Our research shares the same theme in spirit in that we all demonstrate the promise of using online data to forecast future consumer demand. While their studies are typically at the aggregate level using public data, our study shows that an individual firm can actually exploit its private data from click tracking and directly translate it to profit.

The main contributions and findings of the study 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 creates 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).
3. Through a counterfactual study, we show that using the information extracted from the clickstream data can reduce the inventory holding and backordering cost by 3% to 5% in many representative parameter scenarios.
4. To the best of our knowledge, this study is the first in the operations management literature that provides both a model and empirical evidence to demonstrate how the recent clickstream tracking technology can be used to improve operational decisions. Our study aims to stimulate future empirical and theoretical work in this practice- and data-driven field.

The outline of this study 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 section 3, we empirically identify the clickstream variables that are significant for demand forecasting. In section 4, we quantify the operational value of advance demand information from the clickstream data using our model. Section 5 contains the discussion and limitations.

2. A Model of Using Clickstream Data in Inventory Management

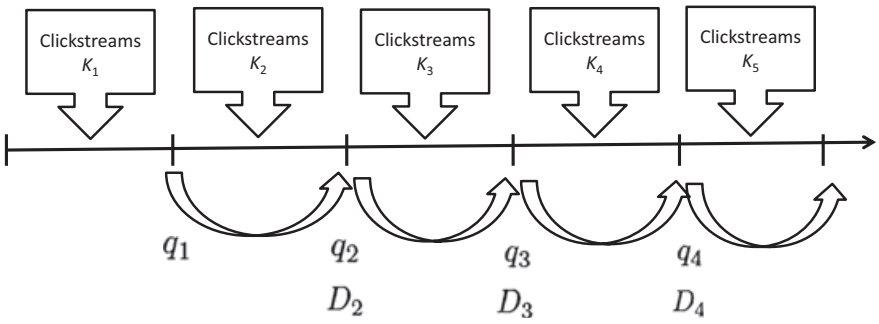
We start by introducing a tactical model of using clickstream data in demand forecasting and inventory management that can serve as a decision support system in practice. This practical model endows classic inventory management with clickstreams as a dynamic flow of advance demand information. In section 3, we will empirically identify relevant model variables. This model will also be our tool for estimating the operational value of clickstream data in section 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 $P_i \equiv f(X_i)$ for $i = 1,2,\ldots$, 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.

Timing. At the beginning of each replenishment period t, the company first satisfies or backorders any realized demand D_t and observes clickstreams of new visitors $\mathbf{K}_t = (K_{1,t}, K_{2,t}, \ldots, K_{J,t})$ that arrived between the beginning of period $t-1$ and the beginning of period t, where J denotes the number of customer classes to be defined below. All the clickstreams observed up to the beginning of period t serve as imperfect advance information of the future demand. Then the company updates its demand forecast, and determines its ordering quantity q_t for input (e.g., a

Figure 1 Description of the Dynamic Programming Model



key "patented part"), which would arrive at the beginning of the next period. 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 heterogenous. Third, in addition to "purchasing" or "never purchasing," a customer now has an additional decision: wait and perhaps 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 homogenous, 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 $j = 1,2,\ldots,J$ and $t' = t + 1, t + 2, \ldots$. 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 $t' = t + 1, t + 2, \ldots$ of never purchasing the product. Clearly, $1 - p_{j,t'} - \gamma_{j,t'} \geq 0$, where the equality always holds in the single-period model but not necessarily in this model for the third reason we pointed out. *Non-buyers* in period t are defined as visitors who will never purchase the product in any future period $t' \geq t$. In a single-period setting, non-buyers are the customers who do not purchase. Hence, non-buyers include what Moe and Fader (2004) define as "hard-core never-buyers." However, in a dynamic setting non-buyers include more than those hard-core never-buyers. It is possible that a customer is interested in purchasing the product initially, say at period t_1, but becomes a non-buyer at period $t_2 > t_1$. Using Moe and Fader (2004)'s term, non-buyers in period t include the "hard-core never-buyers" in all the future periods $t' \geq t$. 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 *indirectly* estimate them in section 4.

We are now ready to describe the system dynamics analytically. Our approach allows for a class-by-class analysis. Recall that $K_{j,t}$ denotes *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*. For brevity, we shall drop the class subscript by writing $K_t \equiv K_{j,t}$ wherever no confusion arises. 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 $Z_{t+1} \equiv Z_{j,t+1}$ 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:

$$Z_{t+1}(z_t; H_t) = z_t + K_{t+1} - D_{t+1}(z_t; H_t) - L_{t+1}(z_t; H_t), \tag{1}$$

which is the previous realized number z_t, plus the number of new potential buyers K_{t+1} from the clickstreams observed in period $t + 1$, minus the demand D_{t+1} and non-buyers L_{t+1}. The non-buyers may not be identifiable from clickstreams, in which case $L_{t+1} = 0$. Typically companies can indirectly estimate the probability that a customer never

purchases from clickstreams. Non-buyers can be identified in cases where the company can obtain some offline information by communicating with the visitors, in which cases the firm should exclude non-buyers from the clickstreams according to Equation (1). Notice that the terms in lower case denote the realizations of the random variables in upper case. In general, Z_{t+1} depends on the entire "history" $H_t \equiv H_{j,t} = (k_1, \ldots, k_t; d_1, \ldots d_t; l_1, \ldots, l_t)$. Let $\mathbf{Z}_t \equiv (Z_{1,t}, Z_{2,t}, Z_{3,t}, \ldots, Z_{J,t})$ and $\mathbf{H}_t \equiv (H_{1,t}, H_{2,t}, H_{3,t}, H_{J,t})$; then the state vector $(\mathbf{Z}_t, \mathbf{H}_t)$ completely describes the system in period t. Clearly, the total demand in period $t+1$: $D_{t+1} = \sum_{j=1}^{J} D_{j,t+1}$.

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 equal 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[2]

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

For $t = 1,2,\ldots,T$, 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).[3] 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 a "clickstreams-dependent" base stock policy, where the optimal order-up-to levels are $y_t^*(\mathbf{z}_t)$. All the parameters required to evaluate the cost saving due to using clickstream data in section 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 $P_i = f(X_i)$ for $i = 1,2,\ldots$ in our model in section 2. To this end, we first discuss our data sets and variable definitions, then specify the general prediction function f as a simple logit or a random-coefficient logit regression equation, and finally show which click variables among X_i are statistically significant.

3.1. Background, Data Source, and Characteristics

The company is in the Midwest of the United States and has some smaller rivals in neighboring states. Consumers can freely shop around and visit websites of multiple similar providers. The website provides comprehensive information to customers; however, due to the customized nature of the product, committing to purchasing is done typically over the phone either through the company directly or through dealers.

The company's website provides the company profile information, product specification information based on different industries, contact information for the company and its dealers, and a webpage where customers can send an email to the company. However, price is not shown on the website and is communicated offline. Customers can acquire information from a few other channels such as phone calls, word of mouth, and brochures from industry conferences. Visiting the website is not a prerequisite for purchasing the product. We do not have an exact percentage of customers that visit the website, as some customers may visit through private computers or their internet service providers that prevent identity

identification.[4] Hence, this study focuses on only those identifiable customers who ever visited the website.

Let us discuss the current inventory management at the company we studied. The company has to keep 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 section 2. The supply 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, as customer demand is satisfied in less than two weeks. (The company can assemble-to-order within two weeks if all required components are available.) The challenge for inventory management is that the supply lead time is much longer than the demand lead time and that backordering customer demand is costly. The intangible adverse effect of the future loss of customer goodwill due to backordering is estimated by managers at around five times of the per unit procurement cost.

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. The company started to track clickstreams from August 2006. The second data set includes both the historical sales data that dates back to March 1998 and recent sales data from August 2006 to November 2008. There are 5185 customers, and 9694 visits in the data.

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 B2B setting where the customers themselves are firms. This has benefits and drawbacks: 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 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 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.

In 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 a summary of the recorded click behavior that includes the time of visits/clicks; cumulative visits (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 US dollars), and the time of ordering.

Before statistical analysis could be started, several preprocessing tasks were executed. First, we cleaned the clickstream data by deleting unidentifiable clicks. 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 even if a company has multiple locations.[5] 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.[6]

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 from August 2006 to November 2008, *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.

Which variables should be used to approximate for customer click behavior? We believe that the answer depends on the context. What we did is to explore all

the commonly used click variables that have been used in the literature (cf. Moe and Fader 2004), for example, cumulative number of visits, visit duration, cumulative and average number of pages, etc. At the same time, we avoid any multicollinearity problem. We also include webpage-specific variables to capture more individual heterogeneity. In our setting, the contact information pages appear informative in terms of predicting purchase propensity.

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. *Cumulative visits*, defined as the cumulative number of visits, 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 a visitor stayed on the website divided by *cumulative visits*. *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 web pages on the firm's website: one category of pages presents product information while the other category shows the contact information if visitors want to contact the company or distributors or if visitors want to become distributors. Intuitively, we expect visits to pages of contact information to be more informative. Indeed, there is a lot of variation in terms of whether these contact-information pages were visited or not, and we use indicator variables to account for this variation. In particular, the variables *contact me*, *contact distributor*, *become distributor*, *reach thanks page* 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 US 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 industry*, *food industry*, *distribution industry*, *manufacturing industry*, *pharmaceutical industry*, *transportation industry*, and *automobile industry* 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. Econometric Model

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"

Table 1 Summary Statistics

Variables	Mean	SD	Min	Max
Ordering behavior				
Order	0.015	0.119	–	–
Order amount ($)	449.26	6249.27	0	286,567.90
Order lead time (days)	89.28	103.31	0	438
General click measures				
Cumulative visits	1.87	2.07	1	30
Average time length (seconds)	229.97	494.51	0.33	10,879.50
Average pages per visit	5.23	9.03	0.23	314.50
Detailed click measures				
Contact me	0.11	0.31	–	–
Contact distributor	0.05	0.21	–	–
Become distributor	0.003	0.057	–	–
Reach thanks page	0.03	0.18	–	–
Historical ordering behavior				
Historical order	0.04	0.19	–	–
Historical order amount ($)	1867	17,791	0	642,375
Industry control variables				
Chemistry industry	0.01	0.11	–	–
Food industry	0.02	0.14	–	–
Distribution industry	0.01	0.09	–	–
Manufacturing industry	0.04	0.19	–	–
Pharmaceutical industry	0.02	0.15	–	–
Transportation industry	0.01	0.10	–	–
Automobile industry	0.01	0.12	–	–

(Sismeiro and Bucklin 2004), logit model (Van den Poel and Buckinx 2005). We refer readers to Hui et al. (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 vs. the alternatives, we actually used their data,[7] and found that the logit model can perform "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 and Buckinx 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, Draganska and Jain (2005, 2006), Train (2003), and Van den Poel and Buckinx (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. 2013 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

$$U_i = \Gamma X_i + \varepsilon_i, \tag{2}$$

where $X_i = [Y_i, Z_i]$ is a vector representing customer i's observed attributes or characteristics. Conceptually and purely for pedagogical purposes and convenience, we can decompose the customer attributes 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 unexplained variation from X_i. Under the assumption that the error terms in Equation (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)

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

The simple logit model has limitations in our setting in that all visitors within each industry share the same coefficients for click variables, although we used demographic variables to take into account visitor heterogeneity.

To incorporate more customer heterogeneity in the prediction function f, we allow *heterogeneity* among the coefficients of click variables even *within* each industry by adopting a random-coefficient logit model.[8] 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 $f(\beta_i \mid \theta^*)$ 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: $T_i(\theta^*) = \int L_i(\beta_i) f(\beta_i \mid \theta^*) d\beta_i$. 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 $P_i(\theta^*) = \int L_{I(i)}(\beta_i) f(\beta_i \mid \theta^*) d\beta_i$. The log-likelihood function is $LL(\theta) = \sum_i \ln P_i(\theta)$. Exact maximum likelihood estimation is impossible, as the integral cannot be calculated analytically. Following Train (2003), we shall approximate the probability through simulation and maximize the *simulated log-likelihood function*.

3.4. Hypothesis Testing

In this subsection, we conduct hypothesis testing to investigate 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 cumulative visits.*

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.*

Now we present our regression results. 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 fail to reject Hypothesis 1, i.e., visitor online click behavior is indeed providing the firm useful information to predict future ordering probabilities.

We find that *cumulative visits* is positively significant at the 1% level. More frequently visiting the website indeed reveals a 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.

Intuitively, how long a customer has been searching may affect or reflect her purchasing propensity. We create a new age factor variable to keep track of how long a customer has been searching: *Searching time length*. This variable measures the time difference between the most recent time of clicking and the first time of clicking (in terms of days) as a proxy for the elapsed time in product searching. As shown in Table 3, it is not statistically significant (p-value = 0.786). This finding may appear surprising. However, it might be explained as follows: *Cumulative visits* measures the *depth* of searching, which is indeed statistically significant. *Searching time length* measures the time *breadth* of searching. A customer may spend a long time in searching without visiting frequently, or she may visit frequently within a short period of time. In our setting, the data suggests that the former behavior tends to suggest this customer is more likely to purchase, i.e., visiting depth rather than visiting time breadth matters more.

More interestingly, from Table 4, not only does *cumulative visits* convey useful information about ordering *probability*, it also provides relevant information about the *timing* of future orders. Indeed, if a

Table 2 Logistic Regression Results (Dependent Variable: Order)

Variable	All customers	New customers	Existing customers
General click measures			
Cumulative visits	0.199**	0.366**	0.160*
	(0.077)	(0.186)	(0.091)
Average time length	0.002	0.004**	−0.002
	(0.001)	(0.002)	(0.001)
Average pages per visit	0.026	0.182	0.004
	(0.063)	(0.160)	(0.098)
Square of average time	−1.18e−06*	−2.50e−06**	−2.87e−07
	(6.75e−07)	(1.12e−06)	(5.18e−07)
Square of average page	−0.0001	−0.007	0.002
	(0.001)	(0.006)	(0.003)
Square of cumulative visits	−0.003	−0.014	−0.002
	(0.003)	(0.013)	(0.004)
Detailed click measures			
Contact me or not	−0.445	−0.186	−0.488
	(0.504)	(0.688)	(0.628)
Contact distributor	1.418**	0.600	1.646*
	(0.610)	(0.792)	(0.858)
Reach thanks page	0.214	0.525	−0.145
	(0.608)	(0.797)	(0.850)
Historical ordering behavior	Yes	Yes	Yes
Industry control variables			
Chemistry industry	0.989	1.714**	
	(0.969)	(0.863)	
Food industry	−0.202	0.770	−0.418
	(0.633)	(1.081)	(0.630)
Distribution industry	−0.822	1.479	
	(1.460)	(0.967)	
Manufacturing industry	0.027	0.245	−0.033
	(0.509)	(1.058)	(0.540)
Pharmaceutical industry	−0.908		−0.711
	(0.604)		(0.766)
Transportation industry	0.617		1.084
	(1.060)		(1.680)
Automobile industry	0.540	1.130	0.352
	(0.633)	(1.043)	(0.651)
Constant	−6.067***	−7.381***	−1.599***
	(0.319)	(0.718)	(0.524)
Pseudo R^2	0.372	0.111	0.143

Standard errors are reported in parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. The number of observations for new customers is 4982, and is 203 for existing customers.

Table 3 Logistic Regression Results with Searching Time Length: Order as the Dependent Variable

Variable	Logit coefficient	Variable	Logit coefficient
General click measures		Industry control variables	
Cumulative visits	0.214***	Chemistry industry	1.136
	(0.076)		(0.818)
Average time length	0.001	Food industry	−0.094
	(0.002)		(0.602)
Average pages per visit	0.055	Manufacturing industry	0.074
	(0.078)		(0.493)
Square of average time	−9.64e−07	Pharmaceutical industry	−0.928
	(1.09e−06)		(0.968)
Square of average page	−0.001	Transportation industry	0.735
	(0.002)		(1.171)
Square of cumulative visits	−0.003	Automobile industry	0.630
	(0.003)		(0.626)
Searching time length	8.71e−06	Constant	−6.160***
	(0.0001)		(0.377)
Detailed click measures			
Contact me or not	−0.439		
	(0.532)		
Contact distributor	1.724***		
	(0.613)		
Reach thanks page	−0.222		
	(0.596)		
Historical ordering behavior			
Historical order	3.545***		
	(0.330)		

Standard errors are reported in parentheses. $^{***}p<0.01$. Pseudo $R^2 = 0.389$.

Table 4 Regression Results: Lead Time as the Dependent Variable

Variable	Tobit coefficient	Variable	Tobit coefficient
General click measures		Industry control variables	
Cumulative visits	−14.760**	Chemistry industry	42.636
	(5.596)		(79.705)
Average time length	0.263	Food industry	57.048
	(0.256)		(47.100)
Average pages per visit	−8.087	Distribution industry	41.306
	(10.454)		(113.371)
Square of average time	−0.0004	Manufacturing industry	−11.167
	(0.0003)		(37.864)
Square of average page	−8.087	Pharmaceutical industry	−31.219
	(10.454)		(81.009)
Square of cumulative visits	0.380*	Transportation industry	−102.875
	(0.208)		(110.032)
Detailed click measures		Automobile industry	−80.148
Contact me or not	33.491		(58.648)
	(51.619)	Constant	155.462***
Contact distributor	−30.564		(34.261)
	(58.316)		
Become a distributor	−13.762		
	(78.517)		
Reach thanks page	−15.690		
	(47.661)		

Standard errors are reported in parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. Pseudo $R^2 = 0.021$.

visitor frequently visits the website, this visitor may be anxious to buy some products in the near future. Hence, her 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 *cumulative visits* and *square of cumulative visits* are significant. Hence, we do not have enough evidence to reject Hypothesis 2.

From Table 5, we can see *cumulative visits*, *square of cumulative visits*, *contact distributor*, and *historical order amount* are significantly and positively associated with *order amount*.[9] Intuitively, more expensive ordering is associated with more frequent visits. In sum, we can use *cumulative visits* to predict both ordering probability, amount, and the timing. These empirical findings confirm that clickstream data provides advance demand information.

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 to 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. Existing customers may have different motivations to visit the website. While new customers visit for information searching, existing customers may visit to get after-sales service. We can also see that *cumulative visits* is just marginally significant (significant at level 10%) from Table 2.

Table 5 Regression Results: Order Amount as the Dependent Variable

Variable	Tobit coefficient	Variable	Tobit coefficient
General click measures		Historical ordering behavior	
Cumulative visits	14,149.12***	Historical order amount	0.35***
	(2499.07)		(0.09)
Average time length	31.26	Industry control variables	
	(41.87)	Chemistry industry	32,449.68
Average pages per visit	3121.84		(23,724.33)
	(2416.55)	Food industry	25,410.78
Square of average time	−0.03		(18,544.97)
	(0.03)	Distribution industry	−15,746.66
Square of average page	−70.55		(39,951.98)
	(77.39)	Manufacturing industry	18,659.26
Square of cumulative visits	−328.02**		(15,424.23)
	(101.40)	Pharmaceutical industry	−26,670.31
Detailed click measures			(33,357.14)
Contact me or not	−20,127.95	Transportation industry	17,553.83
	(16,309.35)		(32,321.93)
Contact distributor	34,150.85*	Automobile industry	28,054.29
	(18,681.09)		(22,635.23)
Become a distributor	15,757.68	Constant	−218,810.60***
	(37,431.07)		(23,605.47)
Reach thanks page	14,755.32		
	(17,403.75)		

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

Table 6 Random-coefficient Logit with Clickstream Coefficients Normally Distributed

Variable	Mean	SD
General click measures		
Cumulative visits	0.428***	0.002
	(0.085)	(0.055)
Average time length	0.001	0.001
	(0.002)	(0.002)
Average pages per visit	0.375**	0.114
	(0.186)	(0.098)
Square of average page	−0.028*	0.007*
	(0.015)	(0.005)
Square of cumulative visits	−0.005	0.0001
	(0.004)	(0.0021)
Detailed click measures		
Contact me or not	−2.911	2.822**
	(1.836)	(1.227)
Contact distributor	0.614	0.043
	(0.899)	(1.211)
Become a distributor	1.529	0.915
	(2.475)	(4.022)
Reach thanks page	0.054	2.476*
	(1.708)	(1.486)
Historical ordering behavior	Yes	Yes
Industry control variables	Yes	Yes

Standard errors are reported in parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

To include more customer heterogeneity, we also estimate the random coefficient logit model. Assuming the coefficients of click variables are normally distributed, we conduct the simulated maximum likelihood estimation using KNITRO-MATLAB 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. 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 *cumulative visits*.

To further examine predictive validity of the clickstream data for demand, we also estimate the logit model using only the randomly selected first half of the data set. Then, we apply the estimated regression equation to the holdout sample (i.e., the second half of the data) and obtain the predicted average purchasing probability (also called conversion rate) 15.61%. Lastly, we compare the predicted average purchasing probability with the actual purchasing probability 14.65%, and get the prediction error in percentage: 6.49% ($\cong (15.61\% - 14.65\%)/14.65\%$). This demonstrates that the predictive power of the clickstream data is fairly good.

We highlight a few findings that are novel compared with those in e-commerce: First, we include more detailed webpage-specific variables that are typically absent in the e-commerce literature (cf. Moe and Fader 2004), and we find that visiting the contact-distributor page or not is useful for predicting future demand. Second, we find differences between new customers and existing customers (e.g., average time length is significant for new customers but not for existing customers). Third, we have the ordering amount information, which is also absent in the literature.

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 section 2 and empirical analysis in section 3.

Which predictors should be tracked? Although the findings here are only for a specific company, the methods do generalize. 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, *cumulative visits*, *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 *cumulative visits* is more than four or not, given that those visitors who visited less than four 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 four 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.[10] 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[11] 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 directly estimate purchasing probabilities p_n and indirectly obtain 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% 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 $p_1 = 0.1046 \times 79.3\% = 0.083$ 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 $t = 2,3,\ldots$, 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, $p_2 = 0.129 \times 0.083 = 0.011$, $p_3 = 0.129 \times p_2 = 0.001$, $p_4 = 0.129 \times 0.001 = 0.0002$, 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 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. As aforementioned, 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 period (i.e., three months) using Figure 1 in section 2. The supply lead time is one period, and the demand lead time is zero.

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 utilized versus when it is not. 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 $Var(D_L) = p_1(1 - p_1)\mathbb{E}Z + p_1^2 Var(Z)$, where Z is the total number of potential buyers expressed in flow Equation (1). Utilizing clickstream data, however, the company can update its demand forecast after observing clickstreams. 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 $Var(D_{L1}) = \mathbb{E}[Var(D_{L1}(Z)) \mid Z] + Var[\mathbb{E}(D_{L1}(Z)) \mid Z] = p_1(1 - p_1)\mathbb{E}Z + p_1^2(1 - p_1)^2\, Var(Z)$. 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 section 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

Table 7 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

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 the parameters within a reasonable neighborhood of the values used earlier. Table 7 summarizes the results and suggests that the cost reduction is typically larger than 3%.[12]

5. Discussion and Limitations

Our primary goal of this study is to show how, and to what extent, clickstream data from non-transactional websites can improve operational forecasting and inventory management. 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. We found that clickstream data can be used to estimate ordering probability, amount, and timing. We also found that advance demand information extracted from the clickstream data can reduce the inventory holding and backordering cost by 3% to 5% in many representative parameter scenarios.

Our study is motivated by practice and is aimed to guide better practice of clickstream tracking in operations management (see also our companion study, Huang and Van Mieghem 2013). 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, 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 that establishes whether clicking causes purchasing or whether it is vice versa. Our data does not allow us to establish such a causal relationship. That requires expensive field experiments for future research. Second, 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.). Third, 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 knowing that the random-coefficient logit model further takes care of the heterogeneity to some degree. Fourth, 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. Fifth, due to analytical tractability and data availability, we cannot incorporate multi-unit demand information for a customer. Hence, this study provides a lower bound for the operational value of the clickstream data. Finally, although our models and methods can be generalized and help build an integrated decision support tool to be 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.

Acknowledgments

We thank the anonymous company for sharing the data with us and Alexandru Rus and Lisa Sun for their assistance in data preprocessing. We are indebted to Zeynep Aksin, Gad Allon, Bariş Ata, Achal Bassamboo, Francis de Véricourt, Sarang Deo, Qi Annabelle Feng, Martin Lariviere, Marcelo Olivares, Özalp Özer, Hyo duk Shin, Che-Lin Su, Anita Tucker, Garrett van Ryzin, the anonymous reviewers, the anonymous senior editor, and Department Editor Panos Kouvelis for many helpful discussions and suggestions that significantly improved the study.

Notes

[1]http://www.ecommercetimes.com/story/19145.html?wlc=1292379670 (Retrieved on December 8, 2012).

[2]This assumption is also made for technical convenience so that we have the unified expression regardless of the sign of the inventory level x. In all other non-terminal period $t < T + 1$, the backorder cost b per unit of time is different from the production cost c. Our assumption is

conservative (given $c < b$) and, hence, provides a lower bound for the operational value of clickstreams.

[3]Notice that our model can be adapted to capture demand from customers who never visited the website. Suppose there is a separate demand $\widetilde{D}_t$ that does not come from the clickstreams in each period t; then our dynamic programming formulation becomes

$$V_t(x, \mathbf{z}_t) = \min_{y \geq x} \{c(y - x) + \beta_0[h\mathbb{E}(y - D_{t+1}(\mathbf{z}_t) - \widetilde{D}_{t+1})^+ + b\mathbb{E}(D_{t+1}(\mathbf{z}_t) + \widetilde{D}_{t+1} - y)^+] + \beta_0\mathbb{E}[V_{t+1}(y - D_{t+1}(\mathbf{z}_t) - \widetilde{D}_{t+1}, \mathbf{Z}_{t+1}(\mathbf{z}_t))]\}.$$

[4]The percentage of buyers who have visited the website among all the buyers is estimated to be around 80.28%. The remaining 19.72% of buyers cannot be found in the cleaned clickstream data.

[5]We also conducted analysis for the sub-group of customers who do not have clickstreams from multiple locations. We found that our qualitative results do not change.

[6]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.

[7]We thank them for generously sharing their data set with us.

[8]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. 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 (Train 2003, Train and Revelt 1998).

[9]If we consider only the customers who did order, then the average order amount is \$31,478.62 and the standard deviation is \$42,227.71. The minimum order amount is \$4124.00. We also did regression analysis for these customers only, and our qualitative finding remains unchanged.

[10]We also did a cluster analysis using the k-means method without *a priori* committing to a belief of the number of classes. Interestingly, the optimal cluster number turns out to be 2, which, to some degree, justifies our heuristic choice in this particular setting.

[11]A Poisson distribution is frequently used in modeling customers arriving at a counter or call center. Suppose there are N customers in the market, and each customer visits the website with probability p. Then the number of visitors to the website would follow the Binomial distribution $\mathbf{B}(N;p)$. In our setting, N is large and p is small (and so the expectation $\mu_K \equiv Np$ is of intermediate magnitude). Then the distribution can be approximated by the Poisson distribution with mean μ_K (by the "Law of Rare Events").

[12]We also implemented the modified dynamic programming model by including the empirical estimate of the demand that does not come from clickstreams and found that the cost reduction is around 2.84%, which is lower than that when only focusing on web visitors. Buyers from non-web-visitors (or unidentifiable web-visitors) tend to dilute the value of using clickstream tracking.

References

Asur, S., B. Huberman. 2010. Predicting the Future with Social Media. arXiv:1003.5699. doi: 10.1016/j.apenergy.2013.03.027

Chen, F. 2001. Market segmentation, advanced demand information, and supply chain performance. *Manuf. Ser. Oper. Manag.* **3**(1): 53–67.

Cheu, R. L., H. Nguyen, T. Magoc, V. Kreinovich. 2009. Logit discrete choice model: A new distribution-free justification. *Soft Comput. – A Fusion Found. Methodol. Applic.* **13**(2): 133–137.

Draganska, M., D. Jain. 2005. Product line length as a competitive tool. *J. Econ. Manag. Strat.* **14**(1): 1–28.

Draganska, M., D. Jain. 2006. Consumer preferences and product-line pricing strategies: An empirical analysis. *Mark. Sci.* **25**(2): 164–147.

Fay, S., D. Mitra, Q. Wang. 2009. Ask or infer? Strategic implications of alternative learning approaches in customization. *Int. J. Res. Mark.* **26**(2): 136–152.

Gallego, G., Ö. Özer. 2001. Integrating replenishment decisions with advance demand information. *Manag. Sci.* **47**(10): 1344–1360.

Gallego, G., Ö. Özer. 2003. Optimal replenishment policies for multiechelon inventory problems under advance demand information. *Manuf. Ser. Oper. Manag.* **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. *Manuf. Ser. Oper. Manag.* **11**(1): 128–143.

Goel, S., J. M. Hofman, S. Lahaie, D. M. Pennock, D. J. Watts. 2010. Predicting consumer behavior with Web search. *Proc. Natl. Acad. Sci., USA*, **107**(41): 17486–17490.

Hariharan, R., P. Zipkin. 1995. Customer-order information, lead-times, and inventories. *Manage. Sci.* **41**(10): 1599–1607.

Huang, T., G. Allon, A. Bassamboo. 2013. Bounded rationality in service systems. *Manuf. Ser. Oper. Manag.* Advance online publication. doi: 10.1287/msom.1120.0417.

Huang, T., J. A. Van Mieghem. 2013. The promise of strategic customer behavior: On the value of click tracking. *Prod. Oper. Manag.* **22**(3): 489–502.

Hui, K., P. S. Fader, E. T. Bradlow. 2009. Path data in marketing: An integrative framework and prospectus for model building. *Mark. Sci.* **28**(2): 320–335.

Johnson, E. J., S. Bellman, G. L. Lohse. 2003. Cognitive lock in and the power law of practice. *J. Mark.* **67**(2): 62–75.

Joo, M., K. Wilbur, Y. Zhu. 2012. Television advertising and online search. Available at SSRN: http://ssrn.com/abstract=1720713.

Levin, N., J. Zahavi. 1998. Continuous predictive modeling: A comparative analysis. *J. Interac. Mark.* **12**(2): 5–22.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers. 2011. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey & Company, New York.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. P. Zarembka (ed.) *Frontiers in Econometrics*. Academic Press, New York, 105–142.

McFadden, D. 2001. Economic choices. *Am. Econ. Rev.* **91**(3): 351–378.

McFadden, D., K. Train. 2000. Mixed MNL models for discrete response. *J. Appl. Economet.* **15**(5): 447–470.

Moe, W., P. S. Fader. 2004. Dynamic conversion behavior at e-commerce sites. *Manage. Sci.* **50**(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. N. Pal, A. Rangawamy. eds. *The Power of One-Leverage Value from Personalizaton Technologies*, eBRC Press, Pennsylvania State University.

Montgomery, A., S. Li, K. Srinivasan, J. C. Liechty. 2004. Modeling online browsing and path analysis using clickstream data. *Mark. Sci.* **23**(4): 579–585.

Özer, Ö. 2011. Inventory management: Information, coordination and rationality. K. Kempf, P. Keskinocak, R. Uzsoy, eds. *Handbook of Production Planning*, Springer, New York, 321–365.

Özer, Ö., W. Wei. 2004. Inventory control with limited capacity and advance demand information. *Oper. Res.* **52**(6): 988–1000.

Porteus, E. 1990. Stochastic inventory theory. *Handbooks in Operations Research and Management Science*, Vol. 2, Elsevier, Amsterdam, 605–652.

Raman, A., M. Fisher. 1996. Reducing the cost of demand uncertainty through accurate response to early sales. *Oper. Res.* **44**(4): 87–99.

Sismeiro, C., R. E. Bucklin. 2004. Modeling purchase behavior at an e-commerce web site: A task completion approach. *J. Mark. Res.* **41**(3): 306–323.

Tan, T., R. Gullu, N. Erkip. 2007. Modeling imperfect advance demand information and analysis of optimal inventory policies. *Eur. J. Oper. Res.* **177**(2): 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. *Rev. Econ. Stat.* **80**(4): 647–657.

Van den Poel, D., W. Buckinx. 2005. Predicting online-purchasing behavior. *Eur. J. Oper. Res.* **166**(2): 557–575.

Wang, T., B. Toktay. 2008. Inventory management with advance demand information and flexible delivery. *Manage. Sci.* **54**(4): 716–732.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1: Additional Support for Adopting the Logit Model