Available online at www.sciencedirect.com

ScienceDirect

Journal of Interactive Marketing 47 (2019) 181–197

JOURNAL OF INTERACTIVE MARKETING

www.elsevier.com/locate/intmar

Invited Article

Analyzing the Browsing Basket: A Latent Interests-Based Segmentation Tool

Nadine Schröder [a,*] & Andreas Falke [a] & Harald Hruschka [a] & Thomas Reutterer [b]

[a] *Department of Marketing, University of Regensburg, Universitätsstraße 31, D-93040 Regensburg, Germany*

[b] *Institute for Service Marketing and Tourism, Vienna University of Economics and Business, Welthandelsplatz 1, Gebäude D2, Eingang A, A-1020 Wien, Austria*

Abstract

The increasing importance of online distribution channels is paralleled by a rising interest in gaining insights into the customer journey to online purchases. In this paper we propose an easy-to-implement two-step procedure that enables online marketing managers to disentangle the complex interrelationships hidden behind observed Internet browsing behavior across websites. Utilizing the procedure allows managers to gain a better understanding why Internet users are visiting their website(s) and how these visits are related to purchases. In the first step, the procedure uncovers latent interests underlying online users' browsing behavior. In the second step, we segment the online users based on their uncovered latent interests. This way, online marketers may understand how segment-specific combinations of latent interests are linked to purchase behavior. We apply the procedure to ComScore clickstream data across 472 websites. We show that there is considerable heterogeneity among online users both regarding online browsing habits, combinations of latent interests, and their conversion into online purchases. For example, some users are interested in apparel and travel service opposed to users who are interested in entertainment tickets. Our empirical analysis confirms that a relatively small fraction of online users realize 70% of online spending. In addition, we detect substantial segment-specific differences of shopping behavior across categories, the most important product categories being apparel as well as food & beverages. Our descriptive perspective comes up with surprising associations among the websites which can be interesting for online marketers.

Keywords: Topic models; Latent Dirichlet allocation; Internet usage behavior; Behavioral segmentation

Introduction

In the wake of the rise of the Internet and the dissemination of user-friendly World Wide Web browser software, online shopping has been steadily growing. In recent years, online retail sales have risen at substantially higher rates than offline sales. In the U.S., online sales reached $453.46 billion in 2017 which represents 13% of total retail sales and 49% of the growth (Zaroban 2018). In Europe the numbers are similar and by 2023, 21% of non-grocery retail sales are expected to be online, up from 13% in 2017 (Forrester Research 2018). With growing penetration rates of tablets and "smart" mobile devices, the Internet is becoming omnipresent, which in turn induces the advent of new digital business models (Brynjolfsson, Hu, and Rahman 2013; Rigby 2011).

Against this background, the Internet continues to play an increasingly important role in information acquisition, product/service evaluation, and price comparisons throughout the purchase funnel prior to online but also offline sales (Bronnenberg, Kim, and Mela 2016; Huang, Lurie, and Mitra 2009; Vuylsteke et al. 2010). On the other hand, sales conversions from commercial website visits remain at very low rates, typically in the lower one-digit percentage range (Moe and Fader 2004; Venkatesh and Agarwal 2006). Consequently, online retailers aim at engaging their visitors in staying longer on their websites and exploring more pages or, in other words, to create "stickiness," which has been shown to be associated with higher profitability (e.g., Bucklin and Sismeiro 2003; Venkatesh

* Corresponding author.
E-mail addresses: nadine.schroeder@ur.de nadine.schroeder@wu.ac.at (N. Schröder), andreas.falke@ur.de (A. Falke), harald.hruschka@ur.de (H. Hruschka), thomas.reutterer@wu.ac.at (T. Reutterer).

https://doi.org/10.1016/j.intmar.2019.05.003
1094-9968

and Agarwal 2006). A number of empirical studies have examined the relationship between website visitation duration and/or the number of page views on purchase incidence (Manchanda et al. 2006; Moe and Fader 2004; Montgomery et al. 2004) or sales (Danaher and Smith 2011).

However, most of this prior research focuses on the browsing and purchase behavior within a given retailer's website. We expand this view by investigating the browsing behavior of online users across different websites. For this purpose, we develop an easy-to-implement segmentation approach that allows uncovering online users' latent interests revealed by web browsing activities and combines the derived interests-based user profiles into segments. Fig. 1 provides a schematic representation of the proposed two-step procedure which is easy to use for online marketing practitioners and also allows Internet marketers to investigate behavioral differences across the derived browsing segments.

The core assumption guiding our proposed segmentation procedure is that online users' observable browsing patterns across multiple commercial websites are driven by some underlying latent interests or online "shopping missions." After preparing and cleaning the available online tracking data (e.g., by selecting websites with at least one purchase, removing websites with extremely high and low numbers of visits; see subsection "*Data*" for further details), in a first step our approach aims at uncovering these latent interests. Our perspective on website visitation patterns is conceptually similar to the way marketing researchers conceive the formation of consumers' "shopping baskets" (e.g., Jacobs, Donkers, and Fok 2016; Manchanda, Ansari, and Gupta 1999). Just as the latter reflect the result of multi-category purchase incidence decisions driven by an individual's context-specific latent shopping preferences, an online user's "browsing basket" reveals combinations of websites she/he considers relevant to satisfy the specific information needs (or latent interests) related to her/his shopping tasks. Methodologically, we employ Latent Dirichlet Allocation (LDA) to infer the latent interests embedded in online users' website visitation patterns. LDA is a method for soft-clustering and commonly used to identify latent topics in large texts (e.g., Blei 2012), which already has seen many applications in the marketing domain (see Reisenbichler and Reutterer 2018 for a recent review). Our LDA application is the first one for mining latent interests from observed browsing behavior at multiple individual websites.

Our approach allows online users to follow a multitude of such latent interests during their browsing history and to combine them with each other. More specifically, in the second step of the proposed procedure we segment the database by clustering the online users' interest-based web browsing profiles. The segmentation derived this way enables online marketers to further examine how specific combinations of latent interests are related to purchase behavior and which product categories benefit the most from specific latent interests. Moreover, knowledge about latent interests that are the drivers behind specific (groups of) online users' "browsing baskets" is relevant for managers who wish to improve their customers' experience by customizing their website to the specific latent interests of their visitors. Many online retailers and non-commercial websites try to accomplish this by personalizing their product recommendations as well as promotional and communication activities based on their users' prior browsing and purchase histories. With the exception of retargeting banner ads, where collaborating websites exchange relevant browsing features (see, e.g., Bleier and Eisenbeiss 2015; Lambrecht and Tucker 2013), however, most of these

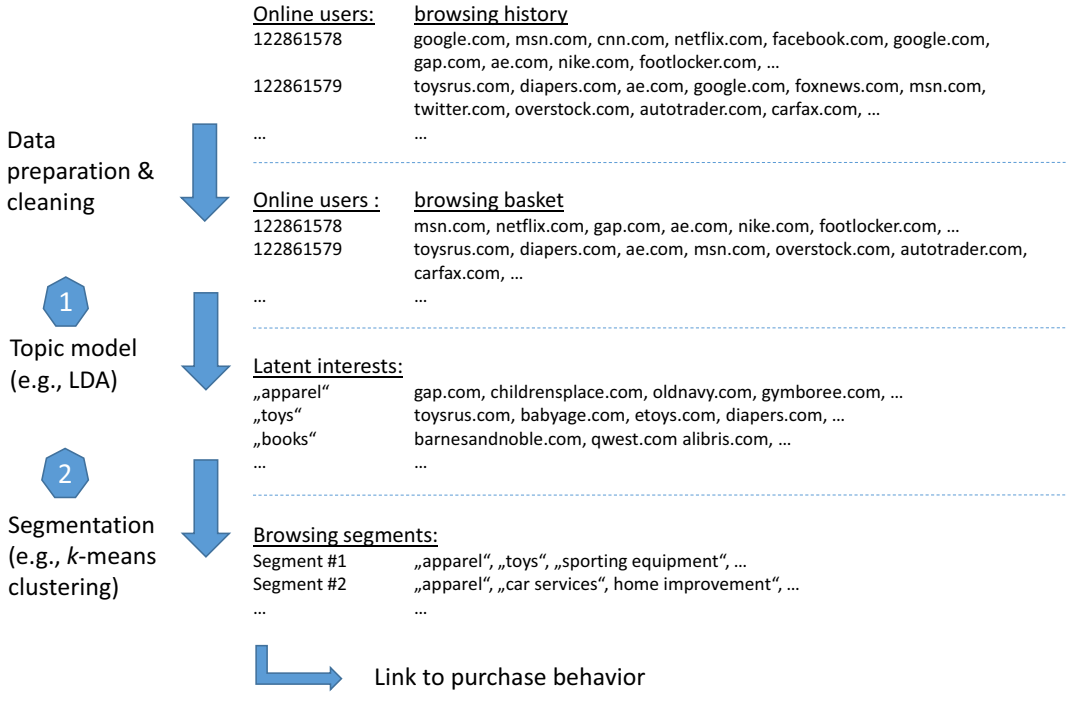

Fig. 1. A two-step segmentation procedure based on latent browsing interests.

customization attempts are currently restricted to the use of internal data available from revisiting users.

The approach we develop and empirically demonstrate in this paper offers online marketers insights into the combinations of latent interests underlying the "browsing baskets" of online users outside a focal firm that can be leveraged to further improve website marketing and customization efforts. For example, website managers could investigate the specific online shopping interests their website is primarily associated with and customize their online marketing activities accordingly. The proposed methodological framework is theoretically well-grounded by combining elements from the shopping basket literature (e.g., Manchanda, Ansari, and Gupta 1999) with the concept of market segmentation (see, e.g., Wedel and Kamakura 2000).

In the next section, we review related literature and position our research against prior studies. Then, we briefly introduce LDA, the data analytic method we adopt to derive latent interests embedded in website visitation patterns. We demonstrate the application of our proposed procedure using the browsing and buying behaviors of a subset of online users participating in the ComScore Web Behavior Panel for 2009 (ComScore 2009). Finally, we discuss implications of our findings and outline further research.

Comparison to Related Studies

The majority of prior contributions, which study online users' browsing behavior, investigate their activities on just one single website by considering sequences of page views. As the focus of our research is different, we review studies which analyze browsing behavior of individual online shoppers or users across multiple websites and product categories. Overall, we found nine studies satisfying these criteria (for an overview see Table 1). Please note that in the following "visit of an online user to a website" is simply called "visit."

Previous related studies use rather different visit related dependent variables, namely the number of page views per visit (Danaher 2007; Danaher, Mullarkey, and Essegaier 2006; Li, Liechty, and Montgomery 2002), visit duration (Danaher, Mullarkey, and Essegaier 2006; Johnson, Bellman, and Lohse 2003), number of websites visited (Johnson et al. 2004), intervisit time (Park and Fader 2004), website choice (Goldfarb 2006), and number of visits (Trusov, Ma, and Jamal 2016). Quite different model types serve to analyze visits. Johnson et al. (2004), Park and Fader (2004) as well as Danaher (2007) apply stochastic models. Johnson, Bellman, and Lohse (2003), Danaher, Mullarkey, and Essegaier (2006), and Mallapragada, Chandukala, and Liu (2016) use random regression models. Moreover, multivariate discrete tobit and nested logit models have been applied by Li, Liechty, and Montgomery (2002) and Goldfarb (2006), respectively. More recently, Trusov, Ma, and Jamal (2016) analyze visits using a topic model.

It is remarkable that among these studies only Johnson, Bellman, and Lohse (2003) and Mallapragada, Chandukala, and Liu (2016) also consider purchase-related measures. Johnson, Bellman, and Lohse (2003) apply a binary logit model for purchases in each of three product categories. Mallapragada, Chandukala, and Liu (2016) use basket value as one of their dependent variables. These authors combine a Tobit-2 model for purchase (binary) and basket value with random regression models for the number of page views and visit duration.

Table 1 shows that the number of product categories considered in these studies vary between 2 and 29. The number of websites vary between 2 and 385. With the exception of Danaher (2007) and Mallapragada, Chandukala, and Liu (2016) previous studies differentiate between website types (e.g.,

Table 1
Comparison to related previous publications.

Publication	Dependent variables	Model type	# of product categories	# of websites or website types	Correlation considered	Time span	Data interval
Li, Liechty, and Montgomery (2002)	# of page views	MVDT	7	7 (Type)	bst	40 mos.	Session
Johnson, Bellman, and Lohse (2003)	visit duration, purchase	RR BL	3	3 (Type)	–	19 mos.	Session
Johnson et al. (2004)	# of websites visited	SM	3	3 (Type)	–	12 mos.	Month
Park and Fader (2004)	intervisit time	SM	2	2 (Type)	wst	7 mos.	Day
Danaher, Mullarkey, and Essegaier (2006)	# of page views, duration	RR	14	7 (Type)	–	1 mo.	Session
Goldfarb (2006)	Website choice	NL	11	2 (Type)	bst	4 mos.	Choice occasion
Danaher (2007)	# of page views	SM	19	15 (45)	bs	1 mo.	Day
Mallapragada, Chandukala, and Liu (2016)	# of page views, duration, basket value	T2 + RR	–	385	–	12 mos.	Session
Trusov, Ma, and Jamal (2016)	# of visits	CTM	29	29 (Type)	bst	12 mos.	Month
This paper	Website visits, purchase frequencies	LDA, *k*-means clusterwise comparisons of product categories	59	472	bs	12 mos.	Week

(Type) indicates that visits to website types are considered, MVDT multivariate discrete tobit, RR random regression, BL binary logit, SM stochastic model, NL nested logit, T2 Tobit 2, CTM correlated topic model, LDA latent Dirichlet allocation, bst between website types, wst within website types, bs between websites.

travel, book, and music websites), which are set a priori and their number varying between 2 and 29.

Danaher (2007) takes correlations of dependent variables between websites into account. Li, Liechty, and Montgomery (2002), Goldfarb (2006), and Trusov, Ma, and Jamal (2016) allow for correlations between websites of different types, Park and Fader (2004) for correlations between websites of the same type. With the exception of Goldfarb (2006) these studies compare to a less complex model which assumes independence of websites. Their results provide evidence that accounting for correlations is important because it leads to better statistical performance. The remaining studies contained in Table 1 ignore the existence of possible correlations among the dependent variables between websites.

The minimum time span of observations in these studies amounts to 1 month, the maximum time span to 40 months. Session, choice occasion, day, and month are alternative basic data intervals to which observations are aggregated.

Against this backdrop of prior research, we summarize the distinguishing aspects of our study as follows:

- We characterize websites as mixtures of latent interests which are based on visit data and determined by LDA and do not introduce fixed website types a priori. The visits that we consider are at the level of individual websites. Our approach differs from Trusov, Ma, and Jamal (2016), who also use a topic model, but look at the number of times an online user visits 29 fixed website types (e.g., services, social media, entertainment). In other words, the authors aggregate visits to the level of website types before analyzing them. As we avoid any classification to a priori fixed website types the latent interests derived by our approach are on a much more fine granular level and thus better reflect the customers' perspectives.
- We allow online users to make multiple choices or pick any choices (see, e.g., Levine 1979) from a high number of different websites. In addition, we take repeated visits of websites during a calendar week into account. These two characteristics show that our approach differs from the work by Goldfarb (2006) which is limited to the choice of one website from a set. As we will explain further in the next section, using LDA gives us the opportunity to analyze such repeated multiple choice data.
- We deal with correlation in a more general manner than the previous studies which take correlation into account with the exception of Danaher (2007) (see Table 1). As mentioned above, these studies either allow correlation between websites of different types precluding correlation of websites of the same type or allow correlation between websites of the same type precluding correlation between websites of different types. We benefit from the fact that LDA is flexible in reproducing correlations by allowing multiple latent interests to be responsible for the websites contained in a visit (Griffiths and Steyvers 2004). Please note that LDA attains this flexibility even though it restricts correlations between latent interest to very low values (Blei, Ng, and Jordan 2003).
- We consider 59 product categories. The maximum number of categories in previous studies amount to 29. Product categories refer to the purchases that panelists make across the various websites and are categorized by the data supplier.
- In contrast to the majority of previous studies, we consider purchase as an additional dependent variable. This way, we recognize that the ultimate goal of many online companies is to increase purchases through their websites (Venkatesh and Agarwal 2006) besides other goals, such as boosting website traffic. We compare yearly purchase frequencies between 59 product categories in different segments of online users. These segments are determined by clustering the importances of topics for each individual online user. Note that only one previous study considers purchases differentiating between (three) different product categories.
- We use a time span of 12 months and aggregate the browsing activities on a weekly time interval. While alternative time references are conceivable as an observation unit as well, our choice should correspond with online shoppers' decision periods, which are typically in the range of a few days or weeks (Johnson et al. 2004; Moe and Fader 2004).
- Finally, by analyzing a total number of 472 unique websites, our research provides a much more comprehensive picture of website visitation behavior across multiple websites than the overwhelming majority of previous studies.

Latent Dirichlet Allocation

In text mining, topic models are often used quite successfully to extract mixtures of topics represented in documents. In this field, words appearing in documents are related to discrete latent variables, which in turn are called topics. Comprehensive descriptions of topic models and typical applications can be found in the text mining literature (see, e.g., Blei 2012; Blei, Ng, and Jordan 2003; Steyvers and Griffiths 2007; Sun, Deng, and Han 2012).

A few publications in which consumer behavior is investigated by topic models have appeared in the marketing literature, which is of relevance in the context of our application. One group of publications focus on text mining of consumers' product reviews. Tirunillai and Tellis (2014) present a text mining study in which they extract latent dimensions of consumer satisfaction by LDA using consumers' online product reviews. Büschken and Allenby (2016) demonstrate that topic models lead to better inference and prediction of consumer ratings compared to simple word-based models. The authors mention that “topic models do not require prior specification of interaction effects and are capable of capturing the pertinent co-occurring words up to the dimensionality of the whole vocabulary” (Büschken and Allenby 2016, p. 954).

The other group of publications analyze consumers' market baskets and benefits from the greater parsimony of topic models compared to, e.g., multivariate choice models, especially if the number of product categories is high. Hruschka (2014) sees topics as latent activities of consumers (linked to, e.g., beverages,

periodicals and cigarettes, or bread and dairy products etc.). He determines topics by both LDA and the correlated topic model (CTM). In principle, the CTM is more flexible by allowing for correlation between topics, but in this study does not perform better than LDA. Hruschka (2014) also demonstrates how product recommendations can be inferred based on conditional purchase probabilities implied by an estimated topic model. Jacobs, Donkers, and Fok (2016) see topics as latent motivations (motivations driven by, e.g., baby related products, or cleaning and personal care). These authors show that topic models attain a better predictive performance compared to collaborative filtering or multinomial models, especially for products which individual consumers have not purchased so far.

The study of Trusov, Ma, and Jamal (2016) investigates website browsing by a topic model. These authors use an extension of the CTM to analyze the number of monthly visits aggregated to 29 website types. Considering both out-of-sample performance and ease of interpretation, Trusov, Ma, and Jamal (2016) choose a CTM with seven topics. We apply LDA, which is the most widespread and computationally efficient topic model to analyze weekly visits of online users to 472 individual websites and thus is well-suited for an easy-to-implement segmentation tool. Contrary to Trusov, Ma, and Jamal (2016) we do not aggregate websites to fixed website types. Our focus on online users is another difference to Trusov, Ma, and Jamal (2016). We only include websites at which at least one purchase has been made during the entire observation period of 1 year and also exclude online users who did not visit any of these included websites.

LDA is based on the assumption that the websites visited by an online user are generated by a mixture of latent interests. Latent interests explain why an online user visits certain websites. All visits share the same latent interests, but their proportions are specific to each visit and randomly drawn from a Dirichlet visit–interest distribution. For each latent interest assigned to a visit this way, a website is chosen randomly from its corresponding distribution. LDA forms latent interests in such a way that websites with higher conditional probabilities for a latent interest frequently co-occur with each other in weekly visits (Crain et al. 2012).

Let I, J, and T denote the number of visits, websites and latent interests, respectively. Random parameters in a (J,T) matrix ϕ and a (T,I) matrix θ indicate the importance of websites for latent interests and the importance of latent interests for visits, respectively. Note that the $t-$th column of ϕ represents the probability of websites conditional on latent interests t and therefore sums up to one.

The probability p_{ij} that visit i contains website j is related to the importance of this website for latent interests and the importance of latent interests for this visit in the following manner (Griffiths and Steyvers 2004):

$$p_{ij} = \sum_{t=1}^{T} \phi_{jt}\theta_{ti}. \quad (1)$$

θ and ϕ are smoothed by Dirichlet hyperparameters α and β. α can be interpreted as prior count of the number of times any latent interest is assigned to a visit, before having observed any website contained in the visit. Low values of α lead to sparse distributions favoring a low number of latent interests. β on the other hand can be seen as prior count of the number of times that websites are sampled from a latent interest before the visit of any website is observed. Each website j of a visit i is linked to latent interests by integer random variables $z_j = 1, \cdots, T$ which give the index of the generated interest. We provide details on the estimation and evaluation of LDA models in Appendix 1.

Using LDA for analyzing such a co-occurrence matrix of visits and websites entails some advantages compared to alternative methods, such as conventional clustering techniques. While the latter could lead to some useful information on latent interests as well, a main weakness of clustering is that visits are typically restricted to be associated with exactly one cluster (or interest). Therefore, clustering ignores the fact that online users may combine several interests in a visit, but assumes instead that all visits assigned to an interest follow one identical interest-specific distribution (Blei, Ng, and Jordan 2003; Steyvers and Griffiths 2007). LDA is free from this problem, because visit probabilities are conceived as convex combination of several interests as shown in Eq. (1). Moreover, in LDA interest probabilities are conditional on the visit containing a website. Also, contrary to cluster analysis, LDA explicitly considers that a topic may arise several times within a visit.

Yet another method, known as latent semantic analysis or latent semantic indexing, is equivalent to principal component analysis applied to the co-occurrence matrix and aims at determining a linear combination of websites to reproduce most of the variance. Continuous principal components are less interpretable and actionable than the discrete interests (topics) discovered by LDA. In addition, latent semantic analysis neglects that online users may visit the same website when pursuing quite different interests (Griffiths, Steyvers, and Tenenbaum 2007).

Empirical Study

In this section we illustrate the empirical application of our two-step segmentation procedure. In doing so, we proceed as follows: After characterizing the analyzed dataset, we present the results obtained from estimating LDA models with different numbers of latent interests (see step 1 in Fig. 1). We select the best performing model using the Bayesian Information Criterion (BIC) and discuss the stability of the derived solution. For the selected model we present and describe the 12 overall most important latent interests. In a second step, we segment online users using k-means clustering with logit-transformed expected frequencies for each online user and each latent interests as input. We interpret each segment with respect to browsing intensity and latent interests, which differentiate between clusters. We also investigate segment differences with respect to purchase behavior. In this context we consider, e.g., the number of purchases, the number of products purchased, and $ sales. In addition, we rank individual product categories

in each segment based on comparisons of average yearly purchase frequencies.

Data

We analyze clickstream data from the ComScore Web Behavior Panel, which were collected from January 1, 2009 to December 31, 2009. Please note that, in the following, we use online users and panelists synonymously. Generally, these data consist of panelists' activities on the Internet (Bucklin and Sismeiro 2009). Our data contain a total number of 173,568,182 visits (before aggregation to a weekly data interval) of 514,440 different websites made by 57,464 panelists. These data also include purchase incidences of individual panelists in 59 different product categories (e.g., apparel, books & magazines, food & beverage, etc.). Despite the huge size of this data set, only 0.2% of the visits (i.e., 411,114 in total) involve a purchase in at least one of these categories.

Because our research emphasizes purchase behavior, we only include websites at which at least one purchase has been made during the entire observation period of 1 year and also exclude panelists who did not visit any of these included websites. From the remaining data set we draw a random sample of 7,500 panelists because of computational limitations. By cross-validating our results we can show that drawing a random sample does not harm the generalizability of our results (see subsection "Model Selection and Stability Analysis of Latent Interests.") In addition, we test if the distribution of two key characteristics of the data set at hand, namely number of visits per panelist and number of visits per panelist and website, is different in the population and the random sample. We do not obtain significant differences as p-values of Kolmogoroff–Smirnoff tests are greater than 0.99.

Furthermore, we use a calendar week as the time frame for studying the composition of the panelists' "browsing baskets." In the following, we thus conceptualize visit as a list of websites accessed by an individual panelist within a specific calendar week.

The resulting browsing baskets comprise a large variety of websites with highly skewed visitation frequencies ranging between 1 and 527,700, with a median number of website visits of 428. Following common data preprocessing practice in text mining (Aggarwal and Zhai 2012, Hoffman et al. 2013, Yogatama et al. 2014), we remove all websites whose number of visits is lower than the 5 percentile or greater than the 95 percentile. The reason behind removing very infrequent websites (words) is, of course, obvious. Removing extremely frequent websites (words) is justified by their inability to discriminate between latent interests (Cho, Fu, and Wu 2017). Using this procedure we retain the overwhelming majority of top-100 U.S. retail websites in 2009 and exclude only three websites, namely bestbuy.com, walmart.com, and amazon.com (Leuenberger 2009). In Appendix 2 we perform some post-hoc analyses on the stability of our results and find that excluding these three websites does not hurt the generalizability of our substantive findings.

Finally, we remove all panelists who never visited any of the remaining 472 websites. Table 2 provides a detailed characterization of the final data set, which we use in our empirical study. On average, we observe 19.1 weekly browsing baskets (visits per panelist) for the 7,235 analyzed comScore panelists. The average number of visits per website amounts to 1,035 and the average number of websites contained in a browsing basket is 3.5 (see Table 2, Part A). Part B of Table 2 contains the purchase frequencies observed for each product category over the one year period. The most important product categories are "online content sales" (mainly consisting of songs from iTunes), followed by "apparel," "shipping services" and "food & beverage" (mainly home delivery services). Notice that the distribution of purchase frequencies across categories is very skewed. For example, the frequency of "online content sales" is about twice as big as the frequency of "apparel." Such unbalanced distributions are typical for online contexts. In the online retailer data analyzed by Jacobs, Donkers, and Fok (2016) the ratio of purchase frequencies between the two most frequent product categories (diapers and laundry detergents) turns out to be similar to what we observe. Therefore, high skewness seems not to be uncommon in online contexts.

Model Selection and Stability Analysis of Latent Interests

In this subsection we discuss how we determine the number of latent interests and examine the stability and interpretability of these latent interests. Deciding on a suitable number of topics is similar to the task of deriving the "optimal" number of clusters when performing model-based clustering. Thus, for the LDA we used the model log-likelihood based BIC measure. We estimate LDA models using blocked Gibbs sampling as implemented in the R package *topic models* (Grün and Hornik 2011). Based on the inspection of log-likelihood traceplots, we discard the first 1,000 iterations for burn-in and calculate estimates from the next 1,000 iterations. α is estimated and β is set to a constant value of 0.1. To avoid local optima, we let the number of latent interests vary between 2 and 110. Due to space limitations, Table 3 only displays BIC values for a selected number of latent interests in ascending order. It becomes obvious from inspection of Table 3 that model fit initially can be clearly improved with an increasing number of topics and the corresponding BIC value reaches its maximum for 86 topics. Thus, we conclude that 86 latent interests best describe the browsing behavior of our sample of households. Note that Trusov, Ma, and Jamal (2016) choose a much smaller number of seven topics. However, the authors consider visits to a much smaller number of only 29 fixed website types, whereas we analyze visits across a diverse set of 472 individual websites.

An alternative approach would be to decide on the number of topics based on managerial judgment and interpretability. Against this background, managers might be keen in reducing the number of topics to a lower number than 86[1]. To get an intuition of how this would affect the topic structure, we

[1] We thank one of the two anonymous reviewers for pointing us to this issue.

Table 2
Descriptive statistics.

Part A: Panelist, visits, and websites

	Panelists	Visits	Websites
	7,235	138,213	472
	Minimum	Average	Maximum
Visits per panelist	1	19.1	53
Visits per website	13	1,035	8,866
Websites per visit	1	3.5	89

Part B: Purchase frequencies of product categories

Apparel	5,491	Shoes	675	Accessories	279
Jewelry & watches	400	Other apparel items	15	Home furniture	311
Home appliances	64	Tools & equipment	64	Kitchen & dining	251
Bed & bath	352	Garden & patio	47	Pet supplies	189
Food & beverage	2,532	Automotive accessories	74	Sport & fitness	268
Health & beauty	2,047	Art & collectibles	131	Tobacco products	52
Baby supplies	201	Other home & living items	395	Books & magazines	2,283
Music	442	Movies & videos	1,183	Desktop computers	27
Laptop computers	87	Handhelds, pdas & portable devices	143	Printers, monitors & peripherals	197
Computer software	202	Other computer supplies	419	Audio & video equipment	114
Cameras & equipment	111	Mobile phones & plans	2,486	Other electronics & supplies	197
Pc video games	5	Console video games	345	Video game consoles & accessories	159
Business machines	3	Office furniture	8	Office supplies	799
Movie tickets	214	Event tickets	488	Air travel	1,350
Hotel reservations	669	Car rental	555	Travel packages	51
Other travel	130	Online content sales	11,706	Online service subscriptions	203
Personals & dating	144	Photo printing services	1,450	Shipping services	2,954
Other services	2,391	Toys & games	481	Arts, crafts & party supplies	564
Flowers	268	Greetings	28	Gift certificates & coupons	74
Other flower & gift items	148	Unclassified	822		

compare the solution of 86 latent interests to two alternative solutions with different numbers of latent interests a, respectively ($a \in \{10, 40\}$). We therefore calculate the dissimilarity score for each of the original 86 latent interests ($t = 1, \cdots, 86$) to each single latent interest from the two alternative solutions ($t_a = 1, \ldots, a$). Our measure for dissimilarities is based on Chaney and Blei (2012):

$$diss_{t,t_a} = \sum_{j \in J} \mid \log\left(1-\phi_{jt}\right) - \log\left(1-\phi_{jt_a}\right) \mid \quad (2)$$

Please note that we use $(1 - \phi_{jt})$ instead of just ϕ_{jt} to emphasize larger importances of a website j for each latent interest (i.e., t, t_a). We say that latent interest t_a matches the original latent interest t if the dissimilarity $diss_{t,\,t_a}$ is lower than the dissimilarities $diss_{t,\,ta}{}'$ for all $t_a' \neq t_a$.

Obviously, reducing the number of latent interests results in having broader latent interests. Let us, as an example, consider one of the latent interests if $a = 10$: If we restrict ϕ_{jt_a} to be at least 0.01, the following websites are associated with this latent interest (with decreasing importances): usps.com, macys.com, intuit.com, walgreens.com, webshots.com, zappos.com, ea.com, cvs.com, shop.com, payless.com, bathandbodyworks.com, landsend.com, shoebuy.com, date.com, llbean.com, ralphlauren.com, christianbook.com, dsw.com, shopnbc.com, shoes.com, and endless.com. It becomes immediately apparent that this latent interest in fact reflects websites from very different areas. To marketers such a latent interest might appear to be rather fuzzy and impractical to derive any substantive insight on the browsing behavior of their respective clientele. Our dissimilarity scores from Eq. (2) support this impression as six out of the 86 original latent interests match this very broad latent interest. For other latent interests from the $a = 10$ solution, we even observe up to 11 matches with the original set of 86 latent interests. Of course, the number of matches reduces with increasing numbers of alternative latent interests. Take as another example a latent interest which belongs to the $a = 40$ solution: One of the exemplary latent interests consists of websites which mainly belong to pharmaceutical (i.e., walgreens.com, cvs.com) and telecommunication websites (i.e., nextel.com). In the original solution with 86 latent interests, two latent interests would match this one, one that solely consists of pharmaceutical websites and another one that consists of telecommunication websites. Of course, each marketer has to decide for herself/himself how concise the

Table 3
Performance of LDA models.

# of latent interests	BIC	# of latent interests	BIC	# of latent interests	BIC
2	−2,315,625	70	−905,822	86	−866,329
10	−1,570,998	80	−877,300	87	−887,605
20	−1,276,433	81	−880,147	88	−872,161
30	−1,120,806	82	−878,696	89	−875,125
40	−1,034,012	83	−872,485	90	−874,241
50	−964,133	84	−871,000	100	−882,393
60	−923,950	85	−877,695	110	−889,689

BIC values rounded to nearest integer.

latent interests should be. As our solution of latent interests appears to be both consistent and face-valid in terms of interpretability (see also Subsection "Analysis of Latent Interests") and statistically supported by the rather parsimonious measure BIC, we thus decide to further explore the solution with 86 latent interests.

To rule out that our results are specific to the chosen random sample of 7,500 panelists and not subject to sampling bias, we cross-validate our findings with 10 additionally extracted random samples of 7,500 panelists each and estimate LDA models with 86 latent interests.[2] To assess the latent interests' (dis-)similarities between the original sample and the 10 additional samples, we again utilize the dissimilarity scores as displayed in Eq. (2). Please note that the number of latent interests for each of the 10 additional samples now always is $a = 86$. In fact, we calculate dissimilarity scores between each of the original 86 latent interests and those from the 86 latent interests derived for the 10 other random samples. We conclude that there is a "match" between a latent interest represented in the original solution with a specific latent interest derived in one of the additional samples for the respective minimum dissimilarity score. Whenever a latent interest from the additional sample received at least one match, we conclude that this latent interest is present across the two examined samples. Using this rationale for the first additional sample we find a match for 81% (=70/86) of latent interests. We proceed in the same manner with the remaining nine additional samples. This way, we find that the vast majority of latent interests show up as well in the 10 additionally extracted samples. In fact, on average 85% (min 81%, max 87%) of the latent interests were recovered across the 10 additional samples. This suggests that the identified latent interests are a decent representation of the browsing behavior of the panelist population.

Table 4
Twelve most important latent interests.

1 =" homeshopping"		2 =" usps.com"		3 =" apparel"	
qvc.com	0.641	usps.com	0.986	gap.com	0.616
hsn.com	0.350			childrensplace.com	0.147
				oldnavy.com	0.129
				gymboree.com	0.047
				bananarepublic.com	0.030
				piperlime.com	0.016
importance = 1,658.8		importance = 1,638.1		importance = 1,635.7	
4 = "home improvement"		5 =" young adults apparel"		6 =" toys"	
lowes.com	0.538	aeropostale.com	0.325	toysrus.com	0.930
homedepot.com	0.412	ae.com	0.295	babyage.com	0.014
acehardware.com	0.036	abercrombie.com	0.139	etoys.com	0.011
		urbanoutfitters.com	0.084	diapers.com	0.011
		delias.com	0.053		
		abercrombiekids.com	0.045		
		alloy.com	0.041		
importance = 1,632.8		importance = 1,631.3		importance = 1,626.8	
7 =" gamespot.com"		8 =" earthlink.net"		9 =" sporting equipment"	
gamespot.com	0.984	earthlink.net	0.990	nike.com	0.280
				eastbay.com	0.236
				footlocker.com	0.213
				finishline.com	0.207
				champssports.com	0.035
importance = 1,625.6		importance = 1,624.1		importance = 1,624.0	
10 =" fedex.com"		11 =" car services"		12 =" overstock.com"	
fedex.com	0.982	autotrader.com	0.797	overstock.com	0.978
		carfax.com	0.152		
		jdate.com	0.027		
importance = 1,623.8		importance = 1,622.9		importance = 1,622.0	

contains all $\phi_{jt} \geq 0.010$ of latent interest t

Analysis of Latent Interests

In the following, we further explore the solution of 86 latent interests. Both the derived latent interests and the websites reflected by these latent interests differ in their contribution to characterize the observed visitation or browsing patterns. Table 4 represents the 12 most important latent interests along with their characterizing websites. Note that we restrict ourselves to a subset of latent interests only for reasons of convenience and space constraints. The remaining latent interests can be characterized in an analogous manner as demonstrated below and more detailed results are available from the authors upon request.

The importance of each latent interest t is measured by its expected frequency, which we obtain by summing θ_{ti} across all visits $i = 1,\ldots, I$. The latent interest with the highest expected frequency is considered to be the most important one. In addition, we indicate importance of a website j for each latent interest t by the estimated ϕ_{jt} value excluding small values $\phi_{jt} < 0.01$.

The most important interest, latent interest no. 1, is related to two websites, i.e., qvc.com and hsn.com. Based on the contents offered by these websites, we label this latent interest "home shopping." On the other hand, latent interest no. 2 is related to only one website satisfying the condition $\phi_{jt} \geq 0.01$. Thus, we name the latent interest after this website, i.e., "usps.com." Both latent interests no. 3 and no. 5 refer to similar websites. Given the relatively broad combination of underlying websites, we label latent interest no. 3 as "apparel." Whereas websites like gap.com and bananarepublic.com are rather classical online apparel stores with mainly adult customers, childrensplace.com and gymboree.com offer apparel for babies and kids. This is in contrast to the websites associated with latent interest no. 5, which we label as "young adults apparel." These websites focus primarily on casual and lifestyle products. Websites belonging to latent interest no. 4 are clearly serving amateurs' needs and we therefore label this latent interest "home improvement." Latent interest no. 6 consists of two different kinds of websites, i.e., toys and layette. However, as website toysrus.com dominates this latent interest ($\phi_{\text{toysrus.com},\ 6} = 0.930$), we label this latent interest "toys." In

[2] We thank the other anonymous reviewer for addressing this issue.

Table 5
Segmentwise browsing behavior.

	Seg. 1	Seg. 2	Seg. 3	Seg. 4	Seg. 5	Seg. 6	Seg. 7
Panelists in %	11	13	15	16	17	17	12
Visits in %	26	23	19	15	10	5	1
Average # of visits per panelist	45.6	34.2	25.3	17.9	11.5	5.9	2
Average # of websites per visit	5.7	3.5	2.8	2.5	2.1	1.9	1.6
Active panelists in % (last two months)	99.7	97.9	95.5	92.1	88.4	79.4	62.9
Latent interest							
"travel tickets and transportation"	L		H		H		
"department store"	H		L		L	L	
"apparel"	H	H		H		L	L
"travel service"			H	H			
"entertainment tickets"	L		H	H	H	H	L
"home shopping"	H			H		L	L
"books"	L						
"apparel & news"	H	L	L	L	L		
"jcpenny.com"	H						L
"travel service (discount)"	L	H	H				

Average importance less than lowest quartile (L), greater than highest quartile (H).

general, several of the obtained latent interests can be seen as refinements of the "shopper's" role presented in Trusov, Ma, and Jamal (2016).

Formation of Panelist Segments

The set of latent interests just derived exhibits two specific properties: The first one is data compression; i.e., the 472 websites are compressed into 86 latent interests which represent combinations of the former. On the other hand, the observed weekly browsing patterns of panelists are generated by mixtures or combinations of multiple latent interests. However, notice that each panelist's specific browsing history can be characterized by "browsing baskets" which are driven by different underlying latent interests. To gain a better understanding on how panelists combine these latent interests over time, we aim at generating segments of panelists and to study their differences with respect to discriminating latent interests and implications for purchase behavior.

In doing so, we first group panelists based on the results of the selected LDA model using *k*-means clustering. For clustering the panelists we calculate the expected frequency f_{ht} of each latent interest t by summing θ_{ti} across all visits of each panelist h and logit-transform it as follows:

$$\log f_{ht} - \log\left(\max_{h'} f_{h't} - f_{ht} + 0.00001\right). \quad (3)$$

This transformation scales the expected frequency of a latent interest t of a panelist h relative to the highest value of this latent interest across all panelists and maps it to the real line. The logit-transformed expected frequencies for 7,235 panelists and 86 latent interests are processed by *k*-means clustering with the number of segments (or clusters) k varying between 2 and 60. We choose the seven segment solution, which reproduces 91.8% of the total sum of squares. Table 5 describes the seven resulting segments. Segments 5 and 6 are the two largest segments each containing 17% of the panelists, while segment 1 is the smallest. By looking at the number of website visits, we obtain quite different results. Segment 1 is largest in this regard and segment 7 the smallest, representing just 1% of overall website visitations.

Segment-Specific Website Browsing Behavior

It turns out that panelists' browsing behavior differs substantially across the derived segments (see Table 5). Members of segment 1 are active almost throughout the whole year, i.e., in 45.6 out of 53 examined calendar weeks. In contrast, panelists in segment 7 seem to browse quite irregularly with an average number of active weeks of just 2. Those households who are active throughout the year also combine more websites in their weekly "browsing baskets"; while segment 1 members visit, on average, 5.7 websites per week, the respective number for segments 6 and 7 are just below 2 websites with the potential of being purchase relevant.

Table 5 also exhibits the percentages of panelists who are active, i.e., who visit any website (including websites which we excluded when estimating the LDA models), in the last 2 months of our observation window. Of course, those panelists with high browsing activity across the year are also active recently (in particular, segments 1 and 2). However, also a substantive fraction of the generally less active online panelists exhibit also browsing activity in the last 2 months. This is a strong indication that segments 6 and 7 do not represent panelists who simply dropped out of the panel but are indeed marked by a substantially lower frequency of visiting websites relevant for panelists. Overall, most website visits are made by the members of segment 1, the fewest visits occur for panelists in segments 7.

Next, we explore whether the derived segments also differ regarding the latent interests characterizing the segment members' online browsing patterns and, if so, which specific latent interests are discriminating between segments the most. To this end, we test each of the 86 latent interests for significant differences in

Table 6
Latent interests differentiating between segments.

"travel tickets and transportation"		"department store"		"apparel"	
cheaptickets.com	0.381	kohls.com	0.878	gap.com	0.616
amtrak.com	0.216	jcpenny.com	0.095	childrensplace.com	0.147
greyhound.com	0.178			oldnavy.com	0.129
res99.com	0.097			gymboree.com	0.047
travelocity.com	0.042			bananarepublic.com	0.030
southwest.com	0.033			piperlime.com	0.016
"travel service"		"entertainment tickets"		"home shopping"	
travelocity.com	0.498	stubhub.com	0.376	qvc.com	0.641
orbitz.com	0.434	ticketmaster.com	0.254	hsn.com	0.350
cheaptickets.com	0.042	ticketsnow.com	0.157		
		tickets.com	0.080		
		ticketliquidator.com	0.055		
		razorgator.com	0.024		
		tickco.com	0.017		
		telecharge.com	0.012		
"books"		"apparel & news"		"jcpenney.com"	
barnesandnoble.com	0.618	wsj.com	0.352	jcpenney.com	0.981
qwest.com	0.140	landsend.com	0.188		
alibris.com	0.061	llbean.com	0.156		
booksamillion.com	0.039	fool.com	0.119		
abebooks.com	0.036	smartbargains.com	0.077		
powells.com	0.031	eddiebauer.com	0.035		
ecampus.com	0.024	onehanesplace.com	0.014		
melaleuca.com	0.023				
"travel service (discount)"					
priceline.com	0.609				
orbitz.com	0.148				
travelocity.com	0.147				
cheaptickets.com	0.076				

contains all $\phi_{jt} \geq 0.010$ of latent interest t

average visitation importances (measured as average expected frequencies) across the seven segments using a series of oneway ANOVAs. Ten latent interests turn out to differentiate significantly between the clusters ($\alpha < 0.05$). For these 10 significant latent interests, Table 5 indicates for each segment whether the average importances are greater (H) or lower (L) than the upper and lower quartiles, respectively. Table 6 gives these latent interests along with their most characterizing websites following the same logic used in Table 4.

Segment 1 members' browsing habits are dominated by the latent interests labeled as "department store," "apparel," "home shopping," "apparel & news," and "jcpenny.com." On the other hand, segment 1 shows low importances for the latent interests "travel tickets and transportation," "entertainment tickets," "books," and "travel service (discount)." Because of this, we refer to these panelists as "Department Store and Home Browsers." Interestingly, the characterizing latent interests observed for these highly active panelists appear to be in contrast to those from segment 3. The latter show particularly strong latent interests in travel-related latent interests and ticketing services, but significantly lower latent interests in websites related to online shopping at department stores and "apparel & news." We hence refer to them as "Travel-focused Leisure Browsers." The online "browsing baskets" generated by segment 2 members seem to share patterns of both segments, which is why we refer to them as "Apparel & Travel Deal Seekers/Browsers." Furthermore, segment 6 also contrasts strongly with the "Department Store and Home Browsers" with significantly lower importances for "department store," "apparel," and "home shopping," but with a high latent interest in "entertainment tickets" and can hence be referred to as "Entertainment Browsers." Panelists from segment 5 show interests in travel service and entertainment tickets which can be summarized as "Leisure Browsers." In addition to that, panelists from segment 4 are also interested in apparel and can therefore be characterized as "Leisure and Home Browsers." The generally least active segment 7 members score low on most of the previously mentioned latent interests. Therefore, we label them as "Occasional Browsers."

In summary, our findings suggest a subset of latent interests which clearly discriminates panelists' Internet browsing behavior. Along this "line of demarcation" we find latent interests related to online shopping activities for product categories offered by department stores including apparel and fashion goods, which shape the browsing behavior of the highly active "Department Store and Home Browsers" representing around 11% of our panel household sample. On the other side, we find a substantial fraction of panel households, in particular "Travel-focused Leisure Browsers" and "Leisure Browsers," that score relatively low on these dimensions but browse the web particularly for travel and ticketing purposes.

Table 7
Segmentwise purchasing behavior.

Segments	Department store and home	Apparel & travel deal seekers	Travel-focused leisure	Leisure and home	Leisure	Entertainment	Occasional
Purchasing panelists in %	81	69	56	44	32	18	5
Visits with purchase in %	11.78	7.61	5.95	4.80	4.3	3.78	3.00
Average # of products bought per purchase	4.35	3.13	3.09	2.18	2.15	2.13	1.89
Average # of products bought per visit	0.256	0.106	0.063	0.029	0.017	0.008	0.002
Total $ sales	530,768	264,745	184,431	100,715	37,502	21,465	2,010
$ sales per panelist	664.29	284.06	175.31	86.90	31.20	17.70	2.29

Table 8
Segmentwise comparison of purchase frequencies between product categories.

Department store and home		Apparel & travel deal seekers		Travel-focused leisure		Leisure and home		Leisure		Entertainment	
Apparel	57	Apparel	56	Apparel	54	Apparel	53	Apparel	51	Air travel	34
Food & beverage	50	Food & beverage	53	Air travel	47	Food & beverage	47	Food & beverage	47	Food & beverage	28
Other services	45	Air travel	44	Food & beverage	43	Air travel	47	Air travel	45	Apparel	25
Health & beauty	43	Photo printing services	40	Photo printing services	43	Photo printing services	43	Hotel reservations	20		
Air travel	43	Other services	37	Event tickets	28	Hotel reservations	35	Event tickets	16		
Shoes	36	Shoes	33	Shoes	26	Event tickets	32				
Photo printing services	36	Event tickets	31	Hotel reservations	25	Shoes	22				
Unclassified	31	Hotel reservations	31	Unclassified	25	Books & magazines	21				
Event tickets	29	Books & magazines	30	Car rental	20	Car rental	17				
Bed & bath	24	Mobile phones & plans	25	Computer software	12	Online content sales	6				
Car rental	23	Car rental	24	Home furniture	5						
Arts, crafts & party supplies	22	Unclassified	19	Bed & bath	2						
Mobile phones & plans	16	Movie tickets	17	Movies & videos	2						
Toys & games	16	Home furniture	16								
Hotel reservations	12	Bed & bath	11								
Accessories	11	Jewelry & watches	10								
Home furniture	11	Other home & living items	10								
Kitchen & dining	11	Toys & games	10								
Movie tickets	8	Computer software	8								
Printers, monitors & peripherals	7	Flowers	5								
Other home & living items	5	Printers, monitors & peripherals	2								
Computer software	5	Arts, crafts & party supplies	2								
Other electronics & supplies	5										
Other flower & gift items	5										
Baby supplies	4										
Other travel	4										
Home appliances	3										
Cameras & equipment	1										

Reading example for apparel and the "Department Store and Home Browsers": for the "Department Store and Home Browsers" the yearly purchase frequency of apparel is significantly higher than the purchase frequencies of 57 other categories.

Segment-Specific Purchasing Behavior

So far, we characterized panelists by their specific combinations of latent interests. Next, we examine how these latent interests are linked to purchasing behavior. Table 7 shows the percentage of panelists making at least one online purchase in 2009. Whereas most "Department Store and Home Browsers" (i.e., 81%) purchase at least once, an almost equally large fraction of online panelists among the "Entertainment Browsers" almost never purchase online in the relevant time period.

The conversion of weekly website visits into purchases is also much higher for the "Department Store and Home Browsers" (with almost 12% of visits) when compared to other segments. In addition, panelists who purchase more frequently also tend to buy more products and spend more money. Again, panelists among the "Department Store and Home Browsers" purchase more products and spend higher amounts online than all the other panelists do. Even though our sample only includes households who visited websites which offer potentially purchase relevant content, about 25% ("Department Store and Home Browsers" and "Apparel & Travel Deal Seekers/Browsers) realize about 70% of overall online sales.

To complement our examination of (combinations of) latent interests underlying website visitation patterns, we next take a closer perspective on the conversion side. To gain a more thorough understanding about which product categories benefit the most from the conversion of website visits into purchases, we systematically analyze segment-specific differences in average numbers of purchases among the 59 product categories we have available in our purchase incidence data. To this end, we conduct $0.5 \times 59 \times 58 = 1{,}711$ pairwise comparisons of category purchases, which implies a Bonferroni corrected significance level of $\alpha = 0.05/1{,}711$ (see, e.g., Jobson 1991). In six out of seven segments, we obtain significantly different category pairs. Note that for segments with very low conversion rates (as given in Table 7) the number of significant differences between product categories decreases considerably. On the two extremes, for the "Occasional Browsers," who have very few purchase incidences, no significant differences between product categories can be observed, while we find among the "Department Store and Home Browsers" the highest number of significant differences.

Table 8 represents, for each segment, a list of product categories ranked in descending order of their respective number of significant comparisons. Note that these lists can be interpreted as rankings of product categories with respect to their importances for online purchases made by the respective segment members. Interestingly, categories "apparel" and "food & beverage" (mostly home delivery services) are always among the top three positions in these segment-specific lists, which implies that these two categories dominate virtually all panelist segments. This is remarkable, because the segments' visitation patterns are driven by rather different combinations of latent interests.

However, the "big picture" that a subset representing about a quarter of panel households ("Department Store and Home Browsers" and "Apparel & Travel Deal Seekers/Browsers") is particularly active (purchases a lot across a wide range of assortment), is confirmed by this category specific view of online purchase activities. In contrast, "Leisure Browsers" and "Entertainment Browsers" show only few product categories with purchase frequencies higher than those of other categories. But there are also some notable differences between the highly active "Department Store and Home Browsers" and "Apparel & Travel Deal Seekers/Browsers" in terms of their purchase behavior. For example, health & beauty attains higher purchase frequencies only for the "Department Store and Home Browsers," whereas books & magazines have high frequencies for "Apparel & Travel Deal Seekers/Browsers." For "Apparel & Travel Deal Seekers/Browsers," hotel reservations clearly play a much more important role as they do in the online shopping baskets of "Department Store and Home Browsers." The contrary applies to categories arts, crafts & party supplies or bed & bath which dominate more other categories among the "Department Store and Home Browsers" as opposed to "Apparel & Travel Deal Seekers/Browsers."

Conclusions

This paper proposes a two-step procedure to assist online marketers in gaining a better understanding of the drivers behind the complex interplay of website visitation streams observed from their users. The presented study is the first application of LDA, an established method for uncovering latent topics in textual databases, to a comprehensive compilation of online users' Internet browsing histories at individual websites. Our approach is based on the assumption that users' observable visitation patterns across a diverse set of online retailers and service providers is driven by a set of underlying latent interests, which are uncovered using LDA. Using one calendar year of clickstream data from a representative subset of 7,235 ComScore web panelists, we show that there is considerable heterogeneity both with respect to online browsing habits, combinations of latent interests, and their conversion into online purchases. In contrast to prior studies (see section "Comparison to Related Studies"), which used similar data to study website browsing behavior, the LDA approach used in our research is capable to reproduce correlations between a large variety of websites and derives latent interests in a purely data driven fashion.

Table 9
Number of latent interests related to website types.

Website types in trusov, ma, and jamal (2016)	# of related latent interests
Automotive	2
Business/Finance	2
Directories/Resources	3
Entertainment	7
Family and youth	1
Gambling	1
Games	8
Health	3
ISP	5
Lifestyles	1
News/Information	5
Promotional/Servers	1
Retail	56
Search/Navigation	2
Services	39
Social media	1
Sports	2
Technology	1
Telecommunications	10
Travel	9

By examining visitation patterns on an extremely fine granular level our approach mimics the customer's perspective of website choices instead of defining website types by the analyst prior to the analysis (which is frequently done in other studies). Table 9 maps the latent interests derived by LDA in our study to 20 of the 29 website types investigated by Trusov, Ma, and Jamal (2016). Note that 14 of these 20 website types are related to multiple (i.e., at least two) latent interests, whereas we observe the two most granular divisions for two website types, namely retail and services with 56 and 39 latent interests, respectively.

Table 10 provides more details on the specific relationships between selected latent interests and these two website types. Our analysis suggests that, as a rule, more detailed latent interests should replace broadly defined website types (e.g., "home shopping," "apparel," "home improvement," "young adults apparel," "toys," "sporting equipment," "overstock.com," "jcpenny.com" instead of retail sites as a whole). Sometimes even individual websites (for example "jcpenny.com" or "ups.com") should replace website types. This illustration shows that the approach demonstrated in this

Table 10
Assigning the eight most important latent interests to retail and service websites.

Retail websites	Service websites
"homeshopping"	"usps.com"
"apparel"	"earthlink.net"
"home improvement"	"fedex.com"
"young adults apparel"	"car services"
"toys"	"southwest.com"
"sporting equipment"	"telecommunications"
"overstock.com"	"dating and body building"
"jcpenny.com"	"car rental and women plus size clothing"

paper is much more fine granular than prior approaches to the problem of structuring website visitation patterns. Furthermore, LDA also allows multiple assignments of a latent interest to several website types (e.g., electronics to both retail websites and service websites). Of course, one loses this kind of information if visits are analyzed only at the aggregate level of website types.

The main findings of our empirical study can be summarized as follows: The weekly aggregated clickstream data representing "browsing baskets" across 472 websites can be adequately compressed into a mixture of 86 latent interests. Online marketers can greatly benefit from this data compression effect resulting from condensing the complex interdependency structures of website visitation patterns into a set of latent interests. In line with the arguments provided by Jacobs, Donkers, and Fok (2016) in a similar context, without such a data compression tool the analyst would be confronted with analyzing an excessively high number of possible co-visitation patterns. More specifically, if we ignore any higher order interactions and the focus is only on pairwise co-visitation frequencies, there would be $J \times (J-1)/2$ cells to evaluate, i.e., 111,156 pairs in the present application with 472 websites (the majority being sparse). Please note that the complexity reduction attained by CTM is much lower compared to the LDA approach which we use. For our situation of 86 topics, CTM would require the estimation of 3,655 (=1/2 * 86 * (86–1)) additional correlation parameters which makes the LDA more in line with our easy-to-use segmentation approach.

Using *k*-means clustering of the panelists' importances devoted to these latent interests, we determined seven segments. These segments are characterized by remarkable differences both in terms of the way they combine various latent interests and in the intensity of their overall online activity. Moreover, these segments also show marked differences in their online purchasing behavior, both in individual product categories and at a more aggregate level. We find that around 25% of panelists ("Department Store and Home Browsers" and "Apparel & Travel Deal Seekers/Browsers") realize 70% of online sales and apparel as well as food & beverages are in all of the examined segments among the dominating product categories. However, we also detected substantial segment-specific differences of shopping behavior across categories.

Managerial Implications

The managerial relevance of the empirical findings resulting from applying our segmentation approach encompasses at least two decision-making scenarios which we briefly illustrate below. Taking the perspectives of an online marketer or digital advertiser, respectively, these scenarios comprise

- the task of adding new promising categories to an online assortment and
- designing user profiles for customizing target marketing actions.

The *first scenario* takes the perspective of a marketer responsible for a particular online shop or service provider who considers to extend the currently offered assortment by introducing new categories. The manager is recommended to carefully examine the (mix of) latent interests which show high importances for the focal website. Based on this, the next step would be to concentrate marketing efforts on segments of online users with high importances for these latent interests and to derive suggestions on which product categories to be adopted or not.

Take as an example the latent interest "travel service" to be important for the focal website. Our empirical study clearly demonstrates that "Travel-focused Leisure Browsers" and "Leisure and Home Browsers" show high importances for the "travel service" interest. These two segments cover about 31% of online panelists, 34% of visits (see Table 5), and almost 11% of visits with a purchase (see Table 7). Based on the purchase conversions for these segments in Table 8, the focal website is recommended to consider the option to offer photo printing services and event tickets, because these categories are very important for the panelists belonging to these two segments and both product categories match well with travel projects. In contrast, most of the other categories, e.g., arts, crafts & party supplies, mobile phones & plans, toys & games, accessories or bed & bath, should be avoided. In addition, Table 5 also shows that "Travel-focused Leisure Browsers" tend to be more interested in travel activities compared to "Leisure and Home Browsers" since they have high importances not only for the latent interest "travel service" but also for latent interests "travel tickets and transportation" and "travel service (discount)." If a manager is considering to use price discounts for their offerings she or he can expect "Travel-focused Leisure Browsers" to be more responsive. As another example, let us focus on the latent interest "apparel & news" to be important for a particular website. This latent interest has a high importance for "Department Store and Home Browsers," which covers about 11% of panelists, 26% of visits and 12% of visits with a purchase. Here, categories like health & beauty, arts, crafts & party supplies are recommended to be considered as add-on categories, but not category books & magazines. These two simple examples illustrate that the empirical findings derived from using the framework proposed in our study can assist online marketers in their considerations of potentially interesting categories to be included in their existing business model beyond those which are more obviously related to the specific latent interests important to their visitors (as, in the above examples, "travel services" or "apparel & news").

As a *second decision-making scenario*, which can greatly benefit from the findings of our study, consider the situation of online marketers and digital advertisers who need to customize their target marketing actions. The latent interests we derived in our approach to detect the driving forces underlying Internet users' browsing behavior correspond to specific combinations of websites. Knowledge about the composition and the particular combinations of these latent interests in attractive segments of online users can be leveraged in designing user profiles to be targeted by specific marketing actions and/or search engine advertising campaigns.

In the following, we will give two examples on how managers of a particular website might benefit from utilizing

Table 11
Latent interests for dickssportinggoods.com.

"sporting equipment"		Ranks according to cross tabulations
dickssportinggoods.com	0.249	1
drugstore.com	0.225	121
sportsauthority.com	0.151	2
nflshop.com	0.091	53
shop.com	0.060	42
vitacost.com	0.036	337
vitaminshoppe.com	0.033	312
modells.com	0.031	59
golfsmith.com	0.027	74
tennis-warehouse.com	0.021	136
skymall.com	0.012	262
contains all $\phi_{jt} \geq 0.010$ of latent interest t		

knowledge about the positioning of their website in the set of latent interests. Furthermore, we contrast these decisions to those resulting from using simple procedures like cross tabulation of co-visitation patterns. For instance, if we take the position of a manager for gap.com she/he would know from the third most important latent interest "apparel" that potential customers who visited gap.com also have childrensplace.com or oldnavy.com in their weekly browsing baskets (see Table 4). Hence, these websites, which are tailored to specific needs such as children's and family apparel, might be good candidates to include into targeting campaigns for gap.com. On the other hand, if the manager looks at cross tabulations by conditioning on visits to gap.com, macys.com or jcpenney.com would result as top candidates since these two websites are ranked highest in terms of co-visit probabilities. Both macys.com and jcpenney.com are intuitive as well, however targeting visitors from these websites might not be as efficient since they offer a much broader product assortment.

As a second example, consider dickssportinggoods.com, which is the online channel of Pennsylvania based "Dick's Sporting Goods" retail company. dickssportinggoods.com is represented with an importance value of $\phi = 0.249$ in one latent interest (see Table 11) which we label as "sporting equipment." An online manager of dickssportinggoods.com who bases her/his targeting strategy on websites with high co-visiting probabilities would choose sportsauthority.com as a possible candidate. However, if she/he based the decision on the latent interest instead, a different decision would emerge. The second largest importance in this latent interest is assigned to drugstore.com, which in 2009 had been a US-based internet retailer in health and beauty care products (https://en.wikipedia.org/wiki/Drugstore.com). Although this combination might be surprising at first sight (as especially beauty care products might be associated with a different target group than sporting goods), the fact that drugstore.com also offered health related products, such as vitamins or dietary supplements, might be of "latent" interest to a target group interested in sporting goods. Thus, adopting the approach advocated in our study to their specific businesses can help managers in developing their "out of the box thinking" when designing effective campaigns. Of course, we do not argue that managers should ignore simple tools like cross tabulations to support these managerial tasks but only relying on them would not have identified drugstore.com as a possible candidate since it has a very low rank (121) in terms of co-visit probabilities to dickssportinggoods.com.

Limitations and Avenues for Further Research

Despite its usefulness for improving online shops and services as well as customizing online marketing campaigns as mentioned above, the approach we presented in this paper also faces some limitations which offer opportunities for future research efforts. In this paper, we use only online browsing and purchasing data and thus ignore potential effects from other more traditional media. In addition, our data set lacks typical marketing variables such as advertising or price changes. We also leave out relationships between browsing and purchases or sales in offline distribution channels.

Of course, from an academic perspective there are many potential avenues to further sophisticate our proposed procedure. In this paper we pursue a two-step approach which is easy to implement for practitioners. We start with a topic model which provides discrete latent variables. In the second step, we obtain clusters of panelists based on these latent variables. One possible path would be to develop and apply a topic model which integrates these two steps by also taking heterogeneity of panelists into account. The community topic model of Li et al. (2012) which simultaneousley detects clusters of authors with similar topics might serve as appropriate starting point. As a consequence, the resulting model would become more complex than our two-step approach and not as easy to implement for online marketers. Another possibility consists in allowing latent variables (interests) to evolve over time. For such an extension, the time-varying information need be included in a topic model. A dynamic topic model as proposed by Blei and Lafferty (2006) to analyze changes of topics over time might be a promising starting point towards this direction. However, such an extension also requires more data spanning over several years. Finally, choosing an LDA model to derive latent interests often involves deleting the extremely frequent terms (i.e., websites) due to their inability to differentiate across latent interests. Although we include almost all big and important websites in our analysis, the exclusion of the three most frequented retail websites (bestbuy.com, walmart.com, and amazon.com) is a further limitation of our study as our results might not be fully generalizable. This issue of dominating "outliers" is also well known in other fields, such as market basket analysis where re-weighting is an option to deal with such issues (see, for example, Strehl and Ghosh 2000). Developing a model that can include such extremely frequented websites is a promising research task which we leave out for future research.

Appendix 1

Estimation and Evaluation of LDA Models

We estimate LDA models by blocked Gibbs sampling, i.e., marginalizing out parameters in ϕ and θ as implemented in the R package *topic models* (Grün and Hornik 2011). Blocked

Gibbs sampling determines the posterior distribution over latent variables z_j (the assignment of websites to latent interests), given the observed websites. For each visit i, the Gibbs sampling procedure considers each website j contained in turn, and determines the probability of assigning the current website to each latent interest, conditioned on the latent interest assignments of all other websites. From this conditional distribution a latent interest is sampled and stored as the new latent interest assignment for this website.

We denote this conditional distribution as $P(z_j = t|z_{-j}, -j, -i)$. $z_j = t$ represents the assignment of website j to latent interest t, z_{-j} the latent interest assignments of all other categories, $-j$ and $-i$ are indices of all other websites and all other visits, respectively. This conditional probability equals (Griffiths and Steyvers 2004):

$$P(z_j = t|z_{-j}, -j, -i) = \lambda \frac{max(n_{1jt}-1, 0) + \beta}{max(n_{1jt}-1, 0) + \sum_{j' \neq j} n_{1j't} + J\beta} \frac{max(n_{2ti}-1, 0) + \alpha}{max(n_{2ti}-1, 0) + \sum_{t' \neq t} n_{2ti'} + T\alpha}. \quad (A.1)$$

λ denotes the proportionality constant. Count variables n_{1jt} and n_{2ti} contain the number of times website j is assigned to latent interest t and the number of times latent interest t is assigned to website j, respectively. Terms $max(n_{1jt} - 1, 0)$ and $max(n_{2ti} - 1, 0)$ in Eq. (A.1) show that the current site and the current visit are not relevant for computing this conditional probability.

The left part of Eq. (A.1) equals the probability of website j under latent interest t. Its right part equals the probability of latent interest t under the current distribution of latent interests for visit i. Once a website has been frequently assigned to latent interest t across all visits, it will increase the probability of assigning any instance of that website to latent interest t. At the same time, if latent interest t has been used many times in a visit, it will increase the probability that any website contained in that visit will be assigned to latent interest t. Therefore, websites are assigned to latent interest depending on how likely the website is for a latent interest, as well as on how important a latent interest is in a visit.

Based on count variables n_{1jt} and n_{2ti} posterior estimates of parameters ϕ_{jt} and θ_{ti} can be computed as (Griffiths and Steyvers 2004):

$$\phi_{jt} = \frac{n_{1jt} + \beta}{\sum_{j=1}^{J} n_{1jt} + J\beta} \quad (A.2)$$

$$\theta_{ti} = \frac{n_{2ti} + \alpha}{\sum_{t=1}^{K} n_{2ti} + T\alpha}. \quad (A.3)$$

We evaluate the performance of LDA models by BIC introduced by Schwarz (1978) which penalizes model complexity:

$$BIC = LL - 0.5 n_p \log(I) \quad \text{with} \quad n_p = T + TJ. \quad (A.4)$$

According to Eq. (A.4) the BIC is based on the log-likelihood LL, the number of visits I and the number of parameters n_p of the topic model. The number of parameters equals the number of latent interests T plus the number of sites J multiplied by T (Blei, Ng, and Jordan 2003). According to Schwarz (1978) the model with the highest BIC is to be preferred.

The log-likelihood LL of a LDA model is computed as follows (Newman et al. 2009):

$$LL = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \left(\sum_{t=1}^{T} \phi_{jt} \theta_{ti} \right) \quad (A.5)$$

n_{ij} indicates how often website j is contained in visit i.

Appendix 2

Post-hoc Analyses on the Exclusion of Selected Popular Websites

In the following, we examine whether the exclusion of the three popular websites bestbuy.com, walmart.com, and amazon.com affects our substantive findings on segment-specific browsing and purchasing behavior. Overall, these three websites account for 5.13% of all website visits and the respective visiting proportions differ only marginally across segments. For instance, we observe the highest proportion for the generally very active segment "Department Store and Home Browsers" (segment 1: 5.43%) and the lowest share for the less active "Entertainment Browsers" (segment 6: 4.70%). An analysis of variance with visiting proportions as the dependent variable confirms that these differences are not significant (p-value = 0.0684) and we thus conclude that exclusion of these websites does not change the overall interpretation of our findings.

Additionally, we investigate whether adding the three websites changes the distribution of purchasing behavior across segments. We base this analysis on the metrics of Table 7. We sum each metric across all seven segments and calculate the share of each segment (see Table A.1). We proceed in the same manner when we include the three websites to the analysis. For example, the share of purchasing panelists is highest among the "Department Store and Home Browsers" both in the original data set (0.267) and in the data set including the three websites (0.254). Overall we find that all shares are more or less identical under the two scenarios. We therefore conclude that the exclusion does not limit the generalizability of our findings.

Table A.1
Segmentwise shares of purchase-related variables.

Segments	Department store and home	Apparel & travel deal seekers	Travel-focused leisure	Leisure and home	Leisure	Entertainment	Occasional
Purchasing panelists (+ 3 websites)	0.267	0.227	0.183	0.143	0.104	0.058	0.018
	(0.254)	(0.222)	(0.183)	(0.145)	(0.110)	(0.063)	(0.023)
Visits with purchase (+ 3 websites)	0.286	0.185	0.144	0.116	0.104	0.092	0.073
	(0.285)	(0.185)	(0.141)	(0.118)	(0.100)	(0.091)	(0.080)
Average # of products bought per purchase (+ 3 websites)	0.230	0.165	0.163	0.115	0.114	0.113	0.100
	(0.221)	(0.164)	(0.159)	(0.123)	(0.117)	(0.115)	(0.101)
Average # of products bought per visit (+ 3 websites)	0.531	0.220	0.132	0.061	0.035	0.017	0.004
	(0.519)	(0.222)	(0.132)	(0.067)	(0.036)	(0.019)	(0.006)
Total $ sales (+ 3 websites)	0.465	0.232	0.162	0.088	0.033	0.019	0.002
	(0.450)	(0.235)	(0.161)	(0.097)	(0.035)	(0.019)	(0.003)
$ sales per panelist (+ 3 websites)	0.461	0.260	0.145	0.076	0.039	0.016	0.002
	(0.446)	(0.263)	(0.145)	(0.083)	(0.042)	(0.017)	(0.004)

References

Aggarwal, Charu C. and ChengXiang Zhai (2012), "A Survey of Text Clustering Algorithms," in *Mining Text Data.* C.C. Aggarwal, C.X. Zhai, editors. New York: Springer, 77–128.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, Jan, 993–1022.

——— and John D. Lafferty (2006), "Dynamic Topic Models," *Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, PA.*

——— (2012), "Probabilistic Topic Models," *Communications of the ACM*, 55, 4, 77–84.

Bleier, Alexander and Maik Eisenbeiss (2015), "Personalized Online Advertising Effectiveness: The Interplay of What, When, and Where," *Marketing Science*, 34, 3, 669–88.

Bronnenberg, Bart J., Jun B. Kim, and Carl F. Mela (2016), "Zooming in on Choice: How Do Consumers Search for Cameras Online?" *Marketing Science*, 35, 5, 693–712.

Brynjolfsson, Erik, Yu Jeffrey Hu, and Mohammad S. Rahman (2013), "Competing in the Age of Omnichannel Retailing," *MIT Sloan Management Review*, 54, 4, 23–9.

Bucklin, Randolph E. and Catarina Sismeiro (2003), "A Model of Web Site Browsing Behavior Estimated on Clickstream Data," *Journal of Marketing Research*, 40, 3, 249–67.

——— and ——— (2009), "Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing," *Journal of Interactive Marketing*, 23, 1, 35–48.

Büschken, Joachim and Greg M. Allenby (2016), "Sentence-Based Text Analysis for Customer Reviews," *Marketing Science*, 35, 6, 953–75.

Chaney, Allison J.B. and David M. Blei (2012), "Visualizing Topic Models," *Proceedings of the Sixth Internaional AAAI Conference on WEblogs and Social Media*.

Cho, Yung-Jan, Pei-Wen Fu, and Chi-Cheng Wu (2017), "Popular Research Topics in Marketing Journals, 1995–2014," *Journal of Interactive Marketing*, 40, November 2017, 52–72.

ComScore (2009), *Web Behavior Disaggregated: Dataset (University of Pennsylvania Wharton Research Data Services)* . dataframe.

Crain, Steven P., Ke Zhou, Shuang-Hong Yang, and Hongyuan Zha (2012), "Dimensionality Reduction and Topic Modeling. From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond," in *Mining Text Data.* C.C. Aggarwal, C.X. Zhai, editors. New York: Springer, 129–61.

Danaher, Peter J., Guy W. Mullarkey, and Skander Essegaier (2006), "Factors Affecting Web Site Visit Duration: A Cross-Domain Analysis," *Journal of Marketing Research*, 42, 2, 182–94.

——— (2007), "Modeling Page Views Across Multiple Websites with an Application to Internet Reach and Frequency Prediction," *Marketing Science*, 26, 3, 422–37.

——— and Michael S. Smith (2011), "Modeling Multivariate Distributions Using Copulas: Applications in Marketing," *Marketing Science*, 30, 1, 4–21.

Forrester Research (2018), *Forrester Analytics: Online Retail Forecast, 2018 to 2023 (Western Europe)*. Cambridge, MA: Forrester Research, Inc.

Goldfarb, Avi (2006), "State Dependence at Internet Portals," *Journal of Economics & Management Strategy*, 15, 2, 317–52.

Griffiths, Thomas L. and Mark Steyvers (2004), "Finding Scientific Topics," *Proceedings of the National Academy of Sciences*, 101, Suppl. 1, 5228–35.

———, ———, and Joshua B. Tenenbaum (2007), "Topics in Semantic Representation," *Psychological Review*, 114, 2, 211–44.

Grün, Bettina and Kurt Hornik (2011), "Topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software*, 40, 13, 1–30.

Hoffman, Matthew D., David M. Blei, Chong Wang, and John Paisley (2013), "Stochastic Variational Inference," *Journal of Machine Learning Research*, 14, 1303–47 May.

Hruschka, Harald (2014), "Linking Multi-Category Purchases to Latent Activities of Shoppers: Analysing Market Baskets by Topic Models," *Marketing ZFP*, 36, 4, 268–74.

Huang, Peng, Nicholas H. Lurie, and Sabyasachi Mitra (2009), "Searching for Experience on the Web: An Empirical Examination of Consumer Behavior for Search and Experience Goods," *Journal of Marketing*, 73, 2, 55–69.

Jacobs, Bruno J.D., Bas Donkers, and Dennis Fok (2016), "Model-Based Purchase Predictions for Large Assortments," *Marketing Science*, 35, 3, 389–404.

Jobson, John D. (1991), *Applied Multivariate Data Analysis. Volume I: Regression and Experimental Design*. New York: Springer . Chapter 5.

Johnson, Eric J., Steve Bellman, and Gerald L. Lohse (2003), "Cognitive Lock-In and the Power Law of Practice," *Journal of Marketing*, 67, 2, 62–75.

———, Wendy W. Moe, Peter S. Fader, Steven Bellman, and Gerald L. Lohse (2004), "On the Depth and Dynamics of Online Search Behavior," *Management Science*, 50, 3, 299–308.

Lambrecht, Anja and Catheine Tucker (2013), "When Does Retargeting Work? Information Specificity in Online Advertising," *Journal of Marketing Research*, 50, 5, 561–76.

Leuenberger, Eric (2009), "Top 100 Retail Websites of 2009," Retrieved 23 January 2019 from http://www.zencartoptimization.com/2009/01/12/top-100-retail-websites-of-2009/.

Levine, Joel H. (1979), "Joint-Space Analysis of Pick-Any Data: Analysis of Choices from an Unconstrained Set of Alternatives," *Psychometrika*, 44, 1, 85–92.

Li, Shibo, John C. Liechty, and Alan L. Montgomery (2002), "Modeling Category Viewership of Web Users with Multivariate Count Models," *Working Paper*, Pittsburgh, PA: Carnegie Mellon University.

Li, Daifeng, Ying Ding, Xin Shuai, Johan Bollen, Jie Tang, Shanshan Chen, Jiayi Zhu, and Guilherme Rocha (2012), "Adding Community and Dynamic to Topic Models," *Journal of Infometrics*, 6, 2, 237–53.Mallapragada, Girish, Sandeep R. Chandukala, and Qing Liu (2016), "Exploring the Effects of "What" (Product) and "Where" (Website) Characteristics on Online Shopping Behavior," *Journal of Marketing*, 80, 2, 21–38.

Manchanda, Puneet, Asim Ansari, and Sunil Gupta (1999), "The "Shopping Basket": A Model for Multicategory Purchase Incidence Decisions," *Marketing Science*, 18, 2, 95–114.

———, Jean-Pierre Dubé, Khim Yong Goh, and Pradeep K. Chintagunta (2006), "The Effect of Banner Advertising on Internet Purchasing," *Journal of Marketing Research*, 43, February, 98–108.

Moe, Wendy W. and Peter S. Fader (2004), "Dynamic Conversion Behavior at E-Commerce Sites," *Management Science*, 50, 3, 326–35.

Montgomery, Alan L., Shibo Li, Kannan Srinivasan, and John C. Liechty (2004), "Modeling Online Browsing and Path Analysis Using Clickstream Data," *Marketing Science*, 23, 4, 579–85.

Newman, David, Arthur Asuncion, Padhraic Smyth, and Max Welling (2009), "Distributed Algorithms for Topic Models," *Journal of Machine Learning Research*, 10, 1801–28.

Park, Young-Hoon and Peter S. Fader (2004), "Modeling Browsing Behavior at Multiple Websites," *Marketing Science*, 23, 3, 280–303.

Reisenbichler, Martin and Thomas Reutterer (2018), "Topic Modeling in Marketing: Recent Advances and Research Opportunities," *Journal of Business Economics*, 1–30.

Rigby, Darrell K. (2011), "The Future of Shopping," *Harvard Business Review*, 89, 12, 65–76.

Schwarz, Gideon (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 2, 461–4.

Steyvers, Mark and Tom Griffiths (2007), "Probabilistic Topic Models," in *Handbook of Latent Semantic Analysis.* T.K. Landauer, D.S. McNamara, S. Dennis, W. Kintsch, editors. Mahwah, NJ: Lawrence Erlbaum, 424–40.

Strehl, Alexander and Joydeep Ghosh (2000), "Value-Based Customer Grouping from Large Retail Data-Sets," *Proceedings SPIE 4057, Data Mining and Knowledge Discovery: Theory, Tools, and Technology II*.

Sun, Yizhou, Hongbo Deng, and Jiawei Han (2012), "Probabilistic Models for Text Mining," in *Mining Text Data.* C.C. Aggarwal, C.X. Zhai, editors. New York: Springer, 129–61.

Tirunillai, Seshadri and Gerard J. Tellis (2014), "Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation," *Journal of Marketing Research*, 51, 4, 463–79.

Trusov, Michael, Liye Ma, and Zainab Jamal (2016), "Crumbs of the Cookie: User Profiling in Customer-Base Analysis and Behavioral Targeting," *Marketing Science*, 35, 3, 405–26.

Venkatesh, Viswanath and Ritu Agarwal (2006), "Turning Visitors into Customers: A Usability-Centric Perspective on Purchase Behavior in Electronic Channels," *Management Science*, 52, 3, 367–82.

Vuylsteke, Alexander, Zhong Wen, Bart Baesens, and Jonas Poelmans (2010), "Consumers' Search for Information on the Internet: How and Why China Differs from Western Europe," *Journal of Interactive Marketing*, 24, 4, 309–31.

Wedel, Michel and Wagner A. Kamakura (2000), *Market Segmentation: Conceptual and Methodological Foundations , International Series in Quantitative Marketing*, Springer.

Yogatama, Dani, Chong Wang, Bryan R. Routledge, Noah A. Smith, and Eric P. Xing (2014), "Dynamic Language Models for Streaming Text," *Transactions of the Association for Computational Linguistics*, 2, 181–92.

Zaroban, Stefany (2018), "U.S. E-Commerce Sales Grow 16.0% in 2017," Retrieved 23 January 2019 from https://www.digitalcommerce360.com/article/us-ecommerce-sales/.