# Analyzing Browsing and Purchasing Across Multiple Websites Based on Latent Dirichlet Allocation

Nadine Schröder*, Andreas Falke*, Harald Hruschka*, Thomas Reutterer†
* Department of Marketing
University of Regensburg, Regensburg, Germany
Email: see http://www.uni-regensburg.de/wirtschaftswissenschaften/bwl-hruschka/lehrstuhl/index.html
† Institute for Service Marketing and Tourism
WU Vienna University of Economics and Business, Vienna, Austria
Email: thomas.reutterer@wu.ac.at

***Abstract*—The increasing importance of online channels for retailers and service providers is paralleled by a rising interest in gaining insights into the customer journey to online purchases. Most attempts to shed light to this issue are restricted to data available for only few particular sites. Our research focuses on mining online shoppers' website visitation patterns across 472 individual websites. We propose a methodological framework to uncover latent interests which we assume to underlie observable online browsing behavior. Using one year of clickstream data for a random sample of comScore panelists, we show that there is heterogeneity among shoppers regarding online browsing habits, combinations of latent interests, and their conversion into online purchases. Our analysis finds that a relatively small fraction of online shoppers realizes 70% of online spending. In addition, we detect substantial segment-specific differences of shopping behavior with respect to 59 product categories.**

***Keywords–Topic Models;Latent Dirichlet Allocation; Internet Usage and Purchasing Behaviour; Behavioral Segmentation***

## I. Introduction

Although online retail sales have grown at substantially high rates in recent years and the internet continues to play an increasingly important role in information acquisition throughout the purchase funnel prior to sales [11][4] sales conversions remain at very low rates [18][23]. 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 [5][23]. However, most of the research focuses on the browsing and purchase behavior within a given retailer's website. In this research, we expand this view by investigating the browsing behavior of online shoppers across different websites and link this behavior with their purchases in several product categories.

We found nine studies analyzing browsing behavior of individual online shoppers across multiple web sites in the marketing and management science literatures [16][13][14][20][6][8][7][17][22]. Seven of these studies do not look at browsing at individual website, but aggregate websites to site types (e.g., travel, book, or music sites).

Let us summarize the novel aspects of our study against to the previous literature. We do not introduce fixed site types, but characterize individual sites as mixtures of latent interests which are based on site visits. Our approach differs from Trusov et al., who also use a topic model, but look at the number of times a consumer visits 29 fixed website types (e.g., services, social media, entertainment) [22]. In other words, these authors aggregate visits to the level of site types before analyzing them. As we avoid aggregation to fixed site types the latent interests, which we obtain, should be better in line with the perspective of consumers. We allow for correlations between all sites which most previous studies have excluded. We consider 59 product categories. The maximum number of categories in previous studies amounts to 29. In contrast to the majority of previous studies, we consider purchase as an additional dependent variable. We compare yearly purchase frequencies between 59 product categories in different segments of online shoppers. These segments are determined by clustering the importances of topics for each individual panelist. Note that only one previous study considers purchases differentiating between (three) different product categories. Finally, by analyzing a total number of 472 unique sites our research provides a much more comprehensive picture of website visitation behavior across multiple sites than the overwhelming majority of previous studies.

In Section II we present the methodological framework, which we adopt to derive latent interests embedded in online shoppers' website visitation patterns. We employ Latent Dirichlet Allocation (LDA), a commonly used technique in text mining to identify latent topics in large texts, which already has also seen promising applications in marketing. In Section III we explain how we obtain the analyzed data by selecting websites and online users participating in the comScore Web Behavior Panel for 2009. In Section IV we present the results of applying LDA to these data. We also segment online users based on their combinations of latent interests and study how different types of online browsing behavior get converted into purchases in a variety of product categories. In Section V we summarize results and outline possible extensions of our approach.

## II. Latent Dirichlet Allocation

In text mining, topic models are often used quite successfully to extract mixtures of topics represented in documents [3][2]. In the following, we define a visit as a list containing all the sites accessed by an individual online shopper in a calendar week. Such a list contains multiple entries for any site, which a shopper accesses several times during a calendar week. We apply LDA, the most widespread topic model, to our data and interpret topics as latent interests. LDA implies the assumption that the sites visited by a shopper are generated by a mixture of latent interests. Let $I$, $J$ and $T$ denote the number of visits, sites and latent interests, respectively. Probabilities $\phi_{jt}$ and $\theta_{ti}$

indicate the importance of site $j$ for latent interest $t$ and the importance of latent interest $t$ for visit $i$, respectively. Please note that the Dirichlet distribution with hyperparameters $\alpha$ and $\beta$ serves as prior for these probabilities.

Finally, the probability $p_{ij}$ that visit $i$ contains site $j$ is related to the importance of this site for latent interests and the importance of latent interests for this visit in the following manner [9]:

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

We see several advantages of LDA in comparison to traditional cluster analytic methods. LDA simultaneously forms soft clusters of sites and visits. It explicitly takes the sparseness of the data into account (as a rule, most sites are not contained in a visit). LDA also considers multiple accesses of the same site during a visit. LDA does not rely on distance measures. It is based individual visits and does not loose information by aggregating across visits.

## III. Data

We analyze clickstream data from the comScore Web Behavior Panel, which were collected from January 1, 2009 to December 31, 2009. Because our research emphasizes purchase behavior, we only include web sites at which at least one purchase is made in one of the 59 categories during the entire observation period. Furthermore, as mentioned before, we use the calender week as time frame. The resulting visits of panelists comprise a large variety of websites with highly skewed frequencies. Following common data preprocessing practice in text mining, each site whose number of visits is lower than the 5 percentile or greater than the 95 percentile is removed. Aggarwal and Zhai recommend to remove very frequent sites (words), as they are not discriminative between latent interests (topics) [1]. Many empirical studies in text mining adhere to this recommendation [10][24]. In fact, our procedure removes only three sites of the top-100 U.S. retail websites in 2009 [15].

Finally, panelists who never visited any of the remaining 472 web sites are removed. The final data set consists of 138,213 visits made by 7, 235 comScore panelists. Each visit is defined as a list of websites accessed by an individual panelist during a specific calender week. To give an example, a list (qvc.com, hsn.com, gap.com, childrensplace.com, qv.com) indicates that a panelist accesses these website (and qvc.com twice) in the respective week. On average panelists make 19.1 weekly visits. The average number of visits per site amounts to 1,035, the average number of sites per visit is 3.5.

## IV. Major Results

### A. LDA Results

We estimate LDA models using blocked Gibbs sampling. The first 1,000 iterations are discarded for burn-in and estimates are based on the next 1,000 iterations. $\alpha$ is estimated and $\beta$ 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.

We evaluate model performance by the Bayesian information criterion (BIC), which penalizes model complexity [21]:

$$BIC = LL - 0.5\, n_p \log(I) \quad \text{with} \quad n_p = TJ. \tag{2}$$

According to Equation (2) 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 $T$ multiplied by the number of sites $J$. The model with the highest BIC is to be preferred. The log likelihood $LL$ of a LDA model ($n_{ij}$ indicates how often site $j$ is contained in visit $i$) is computed as follows [19]:

$$LL = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \left( \sum_{t=1}^{T} \phi_{jt} \theta_{ti} \right) \tag{3}$$

We obtain the best BIC value for 86 latent interests and conclude that 86 latent interests best describe the browsing behavior of our sample of households. Hence, our 472 websites are compressed into 86 latent interests and browsing patterns of online shoppers are generated by combinations of multiple latent interests.

TABLE I. Performance of LDA models

| # interests | BIC | # interests | BIC | # interests | BIC |
|---|---|---|---|---|---|
| 2 | -2,315,613 | 10 | -1,570,939 | 20 | -1,276,315 |
| 30 | -1,120,628 | 40 | -1,033,775 | 50 | -963,837 |
| 60 | -923,595 | 70 | -905,408 | 80 | -876,826 |
| 81 | -879,668 | 82 | -878,211 | 83 | -871,994 |
| 84 | -870,503 | 85 | -877,192 | 86 | -865,820 |
| 87 | -887,090 | 88 | -871,641 | 89 | -874,598 |
| 90 | -873,709 | 100 | -881,801 | 110 | -889,039 |

BIC values rounded to nearest integer

Both the derived latent interests and the sites reflected by these interests differ in their contribution to characterize the observed visitation or browsing patterns. Table II represents the twelve most important latent interests. The importance of each interest $t$ is measured by its expected frequency, which we obtain by summing $\theta_{ti}$ across all visits $i = 1, \cdots, I$. The interest with the highest expected frequency is considered to be the most important one. In addition, we indicate importance of a site $j$ for each interest $t$ by the estimated $\phi_{jt}$ value excluding small values $\phi_{jt} < 0.01$.

Table II illustrates the six most important latent interests. The most important interest # 1 is related to two sites, i.e., qvc.com and hsn.com. Based on the contents offered by these sites we label this topic “home shopping”. On the other hand, interest # 2 is related to only one site satisfying the condition $\phi_{jk} < 0.01$, namely usps.com. Both interests # 3 and # 5 also refer to similar sites. Given the relatively broach combination of underlying sites, we label interest # 3 as “apparel”. Whereas sites 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 sites associated with interest # 5, which we label as “young adults apparel”. These sites focus primarily on casual and lifestyle products. Sites belonging to interest # 4 are clearly serving amateurs’ needs and we therefore label this interest “home improvement”. Interest # 6 consists of two different kind of sites, i.e. toys and layette. However, as site toysrus.com dominates this interest we label this interest “toys”. The remaining latent interests can be characterized in an analogous manner.

TABLE II. Six most important latent interests

| 1 = “homeshopping” | | 2 = ’“postal service 1” | |
|---|---|---|---|
| qvc.com | .641 | usps.com | .986 |
| hsn.com | .350 | | |
| **3 = “apparel”** | | **4 = “home improvement”** | |
| gap.com | .616 | lowes.com | .538 |
| childrensplace.com | .147 | homedepot.com | .412 |
| oldnavy.com | .129 | acehardware.com | .036 |
| gymboree.com | .047 | | |
| bananarepublic.com | .030 | | |
| piperlime.com | .016 | | |
| **5 = “young adults apparel”** | | **6 = “toys”** | |
| aeropostale.com | .325 | toysrus.com | .930 |
| ae.com | .295 | babyage.com | .014 |
| abercrombie.com | .139 | etoys.com | .011 |
| urbanoutfitters.com | .084 | diapers.com | .011 |
| delias.com | .053 | | |
| abercrombiekids.com | .045 | | |
| alloy.com | .041 | | |

gives sites $j$ with $\phi_{jt} >= .010$ for latent interest $t$

TABLE III. Segmentwise browsing behavior

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| panel ists in % | 11 | 13 | 15 | 16 | 17 | 17 | 12 |
| visits in % | 26 | 23 | 19 | 15 | 10 | 5 | 1 |
| average # visits per panelist | 45.6 | 34.2 | 25.3 | 17.9 | 11.5 | 5.9 | 2 |
| average # sites per visit | 5.7 | 3.5 | 2.8 | 2.5 | 2.1 | 1.9 | 1.6 |
| Interest | | | | | | | |
| travel service | | | H | H | | | |
| department store 1 | H | | L | | L | L | |
| apparel | H | H | | H | | L | L |
| travel | L | | 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 | | |
| department store 2 | H | | | | | | L |
| travel service (discount) | L | H | H | | | | |

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

### B. Segment-Specific Website Browsing Behavior

To gain a better understanding on how online shoppers combine these latent interests over time, we aim at generating segments of panelists and study their differences with respect to discriminating latent interests and implications for purchase behavior. 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 interest $t$ by summing $\theta_{ti}$ across all visits of each panelist $h$ and logit-transform it as follows:

$$\log f_{ht} - \log(\max_{h'} f_{h't} - f_{ht} + 0.00001). \qquad (4)$$

We let the number of segments $k$ vary between 2 and 60 and choose a seven segment solution, which reproduces 91.8% of the total sum of squares. Anyway, based on experience with data sets for similar numbers of respondents we did not expect to obtain more than ten segments.

Table III describes the seven resulting segments. In terms of number of panelists segments 5 and 6 are the two largest segments each containing 17%, 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 one percent of overall website visitations.

It turns out that panelists' browsing behavior differs substantially across the derived segments (see table III). 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 visits; 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.

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 interests are discriminating between segments the most. To this end, we test each of the 86 latent interests for significant differences in average visitation importances (measured as average expected frequencies) across the seven segments using a series of oneway analyses of variance. Ten latent interests turned out to differentiate significantly between the segments ($\alpha < 0.05$). For these ten significant latent interests, table III indicates for each segment whether the average importances are less than the lowest quartile (L) or greater than the highest quartile (H). As an example, consider the interest ”travel service”. It consists of the sites travelocity.com, orbitz.com and cheaptickets.com and is very important for segments 3 and 4. We find 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 segment 1 representing around 11% of our panel household sample. On the other side, we find a substantial fraction of panel households, in particular those gathered in segments 3 or 5, which score relatively low on these dimensions but browse the interned particularly for travel and ticketing purposes.

### C. Segment-Specific Purchasing Behavior

In addition, we examine how latent interests are translated into purchasing behavior. Table IV shows the percentage of panelists making at least one online purchase in 2009. Whereas most panelists in segment 1 purchase at least once, about the same fraction of online panelists in segment 6 never purchases online.

The conversion of weekly website visits into purchases is also much higher for segment 1 (with almost 12% of visits) when compared to other segments. In addition, online shoppers who purchase more frequently also tend to buy more products and spend more money. Again, panelists in segment 1 purchase more products and spend higher amounts online than all the other panelists do. About 25 percent (segment 1 and 2 members) realize about 70 percent of overall online sales.

To gain a more thorough understanding which product categories benefit the most from the conversion of site visits into purchases, we systematically compare differences in average numbers of purchases among 59 product categories in each of the seven discussed segments. To this end, we conduct $0.5 \times 59 \times 58 = 1711$ pairwise comparisons of category purchases, which implies a Bonferroni corrected significance level of $\alpha = 0.05/1830$ [12]. In six out of seven segments, we obtain significantly different category pairs. Note that for

TABLE IV. Segmentwise purchasing behavior

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| purchasing panelists | 81% | 69% | 56% | 44% |
| visits with purchase | 11.78% | 7.61% | 5.95% | 4.80% |
| average # of products bought per purchase | 4.35 | 3.13 | 3.09 | 2.18 |
| average # of products bought per visit | 0.256 | 0.106 | 0.063 | 0.029 |
| total $ sales | 530,768 | 264,745 | 184,431 | 100,715 |
| $ sales per panelist | 664.29 | 284.06 | 175.31 | 86.90 |

| | 5 | 6 | 7 |
|---|---|---|---|
| purchasing panelists | 32% | 18% | 5% |
| visits with purchase | 4.3% | 3.78% | 3.00% |
| average # of products bought per purchase | 2.15 | 2.13 | 1.89 |
| average # of products bought per visit | 0.017 | 0.008 | 0.002 |
| total $ sales | 37,502 | 21,465 | 2,010 |
| $ sales per panelist | 31.20 | 17.70 | 2.29 |

segments with very low conversion rates (as given in table IV) the number of significant differences between product categories decreases considerably. On the two extremes, in segment 7 with very few purchase incidences no significant differences between product categories can be observed, while we find in segment 1 most significant differences.

TABLE V. Segmentwise comparisons of purchase frequencies between product categories

| segment 1 | |
|---|---|
| Apparel | 59 |
| Food & beverage | 52 |
| Other services | 47 |
| Health & beauty | 45 |
| Air travel | 45 |
| Shoes | 38 |
| Photo printing services | 38 |
| Unclassified | 33 |
| Event tickets | 31 |
| Bed & bath | 26 |
| Car rental | 25 |
| Arts, crafts & party supplies | 24 |

| segment 2 | |
|---|---|
| Apparel | 58 |
| Food & beverage | 55 |
| Air travel | 46 |
| Photo printing services | 42 |
| Other services | 39 |
| Shoes | 35 |
| Event tickets | 33 |
| Hotel reservations | 33 |
| Books & magazines | 32 |
| Mobile phones & plans | 27 |
| Car rental | 26 |

| segment 3 | |
|---|---|
| Apparel | 56 |
| Air travel | 49 |
| Food & beverage | 45 |
| Photo printing services | 45 |
| Event tickets | 30 |
| Shoes | 28 |
| Hotel reservations | 27 |
| Unclassified | 27 |
| Car rental | 22 |

| segment 4 | |
|---|---|
| Apparel | 55 |
| Food & beverage | 49 |
| Air travel | 49 |
| Photo printing services | 45 |
| Hotel reservations | 37 |
| Event tickets | 34 |
| Shoes | 24 |
| Books & magazines | 23 |

| segment 5 | |
|---|---|
| Apparel | 53 |
| Food & beverage | 49 |
| Air travel | 47 |
| Hotel reservations | 22 |

| segment 6 | |
|---|---|
| Air travel | 36 |
| Food & beverage | 30 |
| Apparel | 27 |

Contains categories with 20 or more significant comparisons. Reading example for apparel and segment 1: for segment 1 the yearly purchase frequency of apparel is significantly higher than the purchase frequencies of 58 other categories.

Table V 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 are always among the top three positions in these segment-specific lists, which implies that these two categories dominate virtually all online shopper segments.

However, the "big picture" of a subset representing about a quarter of panel households (i.e., segment 1 and 2) being particularly active, purchase a lot, and — in addition — do so across a wide range of assortment is confirmed by this category specific view of online purchase activities. On contrary, segments 5 and 6 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 segments 1 and 2 in terms of their purchase behavior. For example, health & beauty and books & magazines attain higher purchase frequencies only in segments 1 and 2, respectively. For segment 2 members, hotel reservations clearly play a much more important role as they do in the visits of segment 1. The contrary applies to categories arts, crafts & party supplies or bed & bath, which dominate more of the other categories in segment 1 as opposed to segment 2.

## V. Conclusion and Future Work

Weekly clickstream data of panelists across 472 websites can be adequately compressed into a mixture of 86 latent interests. Using $k$-means clustering of the panelists' importances devoted to these latent interests, we determine seven online shopper 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 percent of online shoppers (segments 1 and 2) realize 70 percent of online sales and apparel as well as food & beverage are in all of the examined online shopper segments among the dominating product categories. However, we also detect substantial segment-specific differences of shopping behavior across categories.

The approach presented in this paper also faces some limitations which offer opportunities for future research efforts. Here we pursue a two step approach, starting with a topic model, which provides discrete latent variables. In the second step we obtain clusters of panelists based on the importances of these latent variables for the visits of each panelist. To develop and apply a topic model, which integrates these two steps by also taking heterogeneity of panelist into account constitutes an interesting future research endeavor. Another possibility consists in allowing latent variables (interests) to evolve over time. For such an extension, dynamic effects must be included in a topic model. However, such an extension also requires more data spanning over several years.

## References

[1] C. C. Aggarwal and C. Zhai, A Survey of Text Clustering Algorithms: Springer, New York, 2012, pp.77-128, in Aggarwal, C. C., Zhai, C., Mining Text Data.

[2] D. M. Blei, "Probabilistic Topic Models," in Communications of the ACM, vol. 55 , 2012, pp. 77-84.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, Vol. 3, 2003, pp. 993-1022.

[4] B. Bronnenberg, B. J. Kim, and C. F. Mela, “Zooming in on Choice: How Do Consumers Search for Cameras Online?,” Marketing Science, vol. 35 , 2016, pp. 693-712.

[5] R. E. Bucklin and C. Sismeiro, “A Model of Web Site Browsing Behavior Estimated on Clickstream Data,” Journal of Marketing Research, Vol. 40, 2003, pp. 249-67.

[6] P. J. Danaher, G. W. Mullarkey, and S. Essegaier, “Factors Affecting Web Site Visit Duration: A Cross-Domain Analysis,” Journal of Marketing Research, vol. 42, 2006, pp. 182-194.

[7] P. J. Danaher and M. S. Smith, “Modeling Multivariate Distributions Using Copulas: Applications in Marketing,” Marketing Science, vol. 30, 2011, pp. 4-21.

[8] A. Goldfarb, “State Dependence at Internet Portals,” Journal of Economics & Management Strategy, vol. 15, 2006, pp. 317-352.

[9] T. L. Griffiths and M. Steyvers, Finding Scientific Topics, in: Proceedings of the National Academy of Sciences (Suppl. 1), Vol. 101, 2004, pp. 5228-5235.

[10] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, “Stochastic Variational Inference,” Journal of Machine Learning Research, vol. 14, 2013, pp. 1303-1347.

[11] P. Huang, N. H. Lurie, and S. Mitra, “Searching for Experience on the Web: An Empirical Examination of Consumer Behavior for Search and Experience Goods,” Journal of Marketing, vol. 73, 2009, pp. 55-69.

[12] J. D. Jobson, Applied Multivariate Data Analysis. Volume I: Regression and Experimental Design. Springer, New York, 1991.

[13] E. J. Johnson, S. Bellman, and G. L. Lohse, “Cognitive Lock-In and the Power Law of Practice,” Journal of Marketing, vol. 67, 2003, pp. 62-75.

[14] E. J. Johnson, W. W. Moe, P. S. Fader, S. Bellman, and G. L. Lohse, “On the Depth and Dynamics of Online Search Behavior,” Management Science, vol 50 , 2004, pp. 299-308.

[15] E. Leuenberger, Top 100 Retail Websites of 2009, 2009 (http://www.zencartoptimization.com/2009/01/12/top-100-retail-websites-of-2009/ Accessed 14.12.16).

[16] S. Li, J. C. Liechty, and A. L.Montgomery, Modeling Category Viewership of Web Users with Multivariate Count Models. Working Paper, Carnegie Mellon University, Pittsburgh, PA, 2002.

[17] G. Mallapragada, S. R. Chandukala, and L. Qing, “Exploring the Effects of “What” (Product) and “Where” (Website) Characteristics on Online Shopping Behavior,” Journal of Marketing, vol. 80, 2016, pp. 21-38.

[18] W. W. Moe and P. S. Fader, “Dynamic Conversion Behavior at E-Commerce Sites,” Management Science, vol. 50 , 2004, pp. 326-335.

[19] D. Newman, A. Asuncion, P. Smyth, and M. Welling, “Distributed Algorithms for Topic Models,” Journal of Machine Learning Research, vol. 10, 2009, pp. 1801-28.

[20] Y. H. Park and P. S. Fader, “Modeling Browsing Behavior at Multiple Websites,” Marketing Science, vol. 23, 2004, pp. 280-303.

[21] G. Schwarz, “Estimating the Dimension of a Model,” The Annals of Statistics, vol. 6, 1978, pp. 461-464.

[22] M. Trusov, L. Ma, and Z. Jamal , “Crumbs of the Cookie: User Profiling in Customer-Base Analysis and Behavioral Targeting,” Marketing Science, vol. 35, 2016, pp. 405-426.

[23] V. Venkatesh and R. Agarwal, “Turning Visitors into Customers: A Usability-Centric Perspective on Purchase Behavior in Electronic Channels,” Management Science, vol. 52, 2006, pp. 367-382.

[24] D. Yogatama, C. Wang, B. R. Routledge, N. A. Smith, and E. P. Xing (2014), “Dynamic Language Models for Streaming Text,” Transactions of the Association for Computational Linguistics, vol. 2 , 2014, pp. 181-192.