The usage of large data sets in online consumer behaviour: A bibliometric and computational text-mining-driven analysis of previous research

1. Introduction

In the past two decades, advances in information technology as well as the Internet and digitalisation have transformed our daily lives and shifted the focus of commerce towards online digital environments. This change has also altered the ways consumers behave, for including how they shop and buy products and services (Darley, Blankson, & Luethge, 2010; Kim & Lennon, 2008). In the online environment, consumers' decisions to purchase certain products and services are influenced by variables beyond the actual product or service, such as the design of the website or electronic word-of-mouth (including online reviews and recommendations) (see e.g. Cantallops & Salvi, 2014). Thus, online consumer behaviour has been one of the major research areas in marketing science, and it is the subject of a vast number of studies. For example, information system and marketing scholars have examined environmental site features (Manganari, Siomkos, Rigopoulou, & Vrechopoulous, 2011; Richard & Habibi, 2016), global search features (Hausman & Skiepe, 2009), language options (Hausman & Skiepe, 2009), site design (Ha, Kwon, & Lennon, 2007; Mummalaneni, 2005), site security (Hausman & Skiepe, 2009) and culture (Richard & Habibi, 2016).

In recent years, there has been an increasing trend emphasising the importance of data analytics as well as the use of larger data sets. Data analytics is not new. Indeed, various kinds of data have been collected in one form or another for a very long time. However, technological

improvements (e.g. in storing and transmitting information) have enabled the continuous and ubiquitous collection of data. The ability to store this data is also nearly unlimited, due to ever-larger local storage technologies and the rise of cloud storage solutions. Thus, the practice of collecting data on a large scale has become easy and common. However, the challenge for analytics is to go beyond collecting and storing the data and structure it in such a way that the information it contains is accessible and usable. This is where *big data* or *big data analytics*, the hot topic of the past decade, takes its place. *Big data* is a term that refers to both collecting and analysing large data sets; that is, it is a way to build connections within the data set so that the information it contains becomes usable (see e.g. Ruh, 2012; Press, 2013; Erevelles, Fukawa, & Swayne, 2016).

However, despite the above-mentioned emergence of research on online consumer behaviour and large data sets, according to a review article by Darley et al. (2010), the research has been dominated by studies utilising survey data. Over 70 per cent (37 out of 52) of the studies included in their review were based on survey data, including some experiments (31 per cent; 16/52). Thus, the motivation in this paper is not only to discover how research publications on online consumer behaviour cover big data but also how other kinds of large data sets have developed during the last decade and a half. Using a combination of bibliometric analysis and topic modelling, this paper sets out to answer the following research questions:

- How has the research landscape of online consumer behaviour evolved over time?
- What are the dominant areas of research?
- Which authors and articles are most influential?

- Which publication outlets are influential?
- What kinds of topics have been studied?
- What are the avenues for future research on online consumer behaviour?

This study advances the research of online consumer behaviour in multiple ways. First, using a bibliometric citation analysis, the study contributes to the field by identifying the key publications and authors as well as how certain topics have evolved over time. Second, in addition to the traditional bibliometric analysis, which provides mostly descriptive information about the literature, we also use more advanced topic modelling and text analytics techniques. With this more comprehensive approach, it is possible to dig deeper into the literature and see what kinds of topics can be identified in a data-driven way without, for example, restricting the analysis to a predesigned set of potential topics. This provides both new data-driven insights as well as justifications (or a lack of them) for hypotheses about the prevalent research topics. Thus, it is possible to gain more in-depth information about the existing literature.

This paper is organised as follows. Following this brief introduction, Section 2 presents the concepts of large data sets and online consumer behaviour. Section 3 presents the methodology utilised in the paper. Section 4 presents the findings of the traditional bibliometric analysis as well as the results of the topic modelling. Finally, in Section 5, we discuss the results, present the conclusions and limitations and provide suggestions for further research.

2. Large data sets and online consumer behaviour

It has been argued that consumer behaviour in the online context is not that different from consumer behaviour in traditional settings and that patterns of general consumer behaviour are also applicable to the online context (Lee & Chen, 2010). For example, Koufaris (2002) stated that the traditional and online environments clearly share characteristics, but, within the online environment, consumers have additional unique needs and concerns because of their two-fold roles as both consumers and information system users (Lee & Chen, 2010). Montgomery, Li, Srinivasan and Liechty (2004) claimed that compared to traditional consumers, the online consumer is more convenience-oriented and innovative and seeks more variety. Consequently, online consumer behaviour can be defined as a manifestation of consumers' decision-making and behaviour in the online environment as well as the factors that influence this behaviour when, for example, customers search for information, browse websites, evaluate products or compare information.

In recent years, the average consumer has turned into a constant generator not only of traditional (i.e. structured and transactional) data but also of contemporary, unstructured and behavioural data. The magnitude of this generated data, the rapidity of its generation and its diverse richness impact how it should be treated and interpreted. As a natural consequence, big data has emerged as a topic in the related research literature, although there is a lack of a coherent use of the term (see e.g. Gandomi & Haider, 2015). In the strict sense, big data means data sets that have become too large and complex for conventional statistical measures, thus requiring advanced and unique methods for data storage, management, analysis and visualisation (Chen, Chiang, & Storey, 2012). Or, as Erevelles et al. (2016) put it (see also De Mauro, Greco, Grimaldi, & Ritala, 2018;

Sivarajah, Kamal, Irani, & Weerakkody, 2017), data sets have become larger and more complex because of the unprecedented volume, velocity and variety of primary data available from individual consumers. According to Manyika et al. (2011; see also Erevelles et al., 2016), the underlying methodology for big data analytics is linked to existing disciplines, mainly statistics and computer science. Big data can be defined as high-volume, high-velocity and high-variety data assets (commonly referred to as the three Vs), which demand innovative processing methods so that the resulting information can be used to derive enhanced insights for decision-making (Erevelles et al., 2016; Hofacker, Malthouse, & Sultan, 2016; Lycett, 2013).

Consumer behaviour produces large amounts of certain types of data. One of these is *clickstream data*, defined as an electronic record of Internet usage collected from the server log of a website or by third-party services. It typically contains data regarding individual navigational mouse clicks and clicks on other elements of web pages that a visitor makes when visiting a certain website. This kind of data provides information about the sequence of pages viewed by users as they navigate a website (Bucklin & Sismeiro, 2009; Montgomery et al., 2004; Moe, 2003; Senecal, Kalczynski, & Nantel, 2005). Another type is *text data,* which includes all kinds of natural language text, such as web pages, emails and social media posts (e.g. Zhai, 2017).

A particular use of online consumer data is *path analysis*, which is based on the data about the navigational path a user takes through the website as well as insights based on the analysis of that data. In other words, path data analysis produces information about the sequence of events leading to a decision to purchase. This information typically contains data about the user's goals,

knowledge and interests, which can be used to predict consumer behaviour (Bucklin & Sismeiro, 2009; Montgomery et al., 2004). Meanwhile, *customer analytics* is defined as the overall process by which data regarding customer behaviour are used to help make business decisions. Analytics lies at the junction of data and consumer behaviour; data provide insights regarding customer behaviour, for example, and marketers turn these insights into a market advantage (see e.g. Corrigan, Craciun, & Powell, 2014; Erevelles et al., 2016).

3. Methodology

3.1 Data collection

The initial search for potential articles took place in February 2019. We collected data from the academic database ISI Web of Science (also known as Web of Knowledge) and its Core Collection. According to Fetscherin and Heinrich (2015; see also Fetscherin & Usunier, 2012), the ISI Web of Knowledge is a suitable database when the object is to conduct an interdisciplinary literature review, and many notable bibliometric studies have used this database.

We searched for publications that appeared from January 2000 to December 2018. The year 2000 was chosen as the cut-off year, as it could be considered the first year when the usage of large data sets emerged in the literature. December 2018 marked the most recent date for which complete citation data from the ISI Web of Science were available to us.

To collect comprehensive data, we searched for *journal articles* written *in English* based on the appearance of search terms in the topic field in the Web of Science. The topic field includes the title, abstract and keywords of articles. It also includes so-called 'keyword plus' content, additional relevant but overlooked keywords that were not listed by the authors or publishers but which are based on the expertise of the editors of the Web of Science. We utilised our own three-facet classification for the search terms. The first level was named 'data level' and represented which kind of data was used in the articles. The search terms for this level were 'big data', 'clickstream', 'text data', 'path analysis' and 'customer analytics'. The second level was labelled 'individual level', representing the role of the individual. The search terms 'customer' and 'consumer' were utilised for this level. Finally, the third level represents the context in which the consumer or customer operates. For this level, we included the terms 'purchase', 'buying', 'online discussion', 'information search', 'customer journey', 'online', and 'mobile' in our search. The search was conducted so that search criteria within each level were complementary alternatives (i.e. it was enough that one of the search terms within the level should occur); for this, we utilised the Boolean operator 'OR'. Between the levels, the search criteria had to be fulfilled at the same time (i.e. at least one search term should occur from each of the levels); for this the Boolean operator 'AND' was utilised. The total set of search criteria we used in the Web of Science is presented in Table 1.

INSERT TABLE 1 ABOUT HERE

As a result of the search, we found 498 articles. Based on a manual inspection of these articles, we removed three articles because they were not actually related to large data sets. Eventually, 495 articles were included in the bibliometric analysis. We categorised the articles into eight

different groups including ‘Business and Economics’, ‘Computer Science and Information Science’, ‘Engineering’, ‘Health’, ‘Hospitality’, ‘Environmental Sciences’, ‘Communication’, and ‘Other’ based on the types of journals to which they belonged, where the ‘Other’ group contained literature from, for example, chemistry-, mathematics- and education-related journals. Table 2 lists the number of publications from different fields.

INSERT TABLE 2 ABOUT HERE

3.2 Bibliometric analysis

A bibliometric data analysis was conducted to provide a quantitative analysis of the academic literature. Bibliometric analysis is the statistical analysis of written publications and citation analysis, and it is based on the construction of a citation graph, a network or graph representation of the citations between documents. Many research fields use bibliometric methods to explore the impact of their field, of a set of researchers or of a particular paper (e.g. Nicolaisen, 2010; Fetscherin & Heinrich, 2015; Hajikhani, 2017).

For this study, the bibliometric data analysis was conducted using the Network Analysis Interface for Literature Studies (NAILS) tool (Knutas, Hajikhani, Salminen, & Porras, 2015). The NAILS tool analyses several essential variables from each publication, including the authors, keywords, publication forum, article type and cited articles. After the data are downloaded from Web of Science, the tool removes duplicate records and performs an exploratory data analysis, which identifies the most-cited articles and authors, the most common keywords and the journals with the most publications. The system also extracts the citation network data from the publications. In addition to the exploratory analysis report, the tool further

provides data about citation and author cooperation networks, which can be visualised. This kind of data set about citation connections can be used to calculate the relative influence of publications in the network (Knutas et al., 2015).

In the NAILS tool, several metrics (e.g. PageRank and In-Degree) were calculated based on the citation data. PageRank (Brin & Page, 1998) counts the number and quality of links to a paper to obtain a rough estimate of its importance, while In-Degree provides the number of citations referring to an individual paper based on the citation data records. Utilising these metrics, a report was produced, which provided an abstract and keyword analysis as well as the most productive authors and journals.

3.3 Text mining and topic modelling

Text mining by topic modelling aims to discover topics that occur in a collection of documents in order to explore hidden semantic structures in the body of the texts. In this work, the text mining and topic modelling was conducted using the structural topic model (STM; Roberts, 2014). Although the most common topic modelling technique is Latent Dirichlet Allocation (LDA; Blei, 2012) and an implementation of it is even built into the NAILS tool (Hajikhani, 2017; based on a visualised application of LDA by Sievert and Shirley, 2014), we decided use the more advanced and sophisticated STM method. In short, this is because STM both turned out to be a quantitatively better fitting probabilistic model for the document corpus and because, unlike LDA, it is directly able to model the interaction of covariates like time with the topics; additionally, the specific NAILS implementation of LDA has further restrictions. We next

present the details of STM and its training and then describe the advantages of STM over LDA in more detail in Section 3.3.2.

In order to enrich the text source, we combined text from the document title, keywords and abstract. Note that this process allows the topic model to emphasise terms if they are mentioned in more than one location (e.g. both in the abstract and in the keywords). As said, we used a more advanced text analysis method (STM), which allowed us to obtain more insights by exploring the relationships between latent topics and the covariates we were interested in. The STM model has been widely used in different disciplines, for example, as part of evaluating the influences of censorship in China by analysing blog posts (Roberts, 2014) and as part of analysing open-ended questionnaires to explore public opinions on climate change (Tvinnereim & Fløttum, 2015).

3.3.1 Topic modelling based on the STM

Like LDA, the STM models the documents as arising out of a mixture of underlying topics with different prevalences, and it models the topic content, in turn, as a mixture of word or bigram terms with different prevalences. Different topics essentially represent different groups of frequently co-occurring words or bigram terms. Each word is associated to the different topics with different prevalences (probabilities); the topics can then be summarised by their most prevalent (most probable) words. Expressing the description above in a mathematical way, the model first generates a topic prevalence parameter θ_d for each word containing the probability of $k = 1, \ldots, K$ different topics. Then, for each topic k, the model generates topic content parameter β_k over $v = 1, \ldots, V$ vocabularies. The nth word in the document d, w_n is generated by first drawing a topic label z_n from a multinomial distribution with parameter θ_d and then drawing a

word from the corresponding word from another multinomial distribution with parameter β_{z_n}. However, unlike the basic LDA model, the STM model assumes that the topic prevalence and content can depend on some document-level covariates. That is, in STM, θ_d is generated in a regression-like procedure, drawn from a Logistic-Normal distribution:

$$\theta_d \sim Logistic - Normal(\Gamma' x_d, \Sigma)$$

where x_d are the document-level covariates, Γ is the matrix of coefficients and Σ is a hyper-parameter. Further, the topic content β has a tensor structure and consists of

$$\beta_{d,k,v} = \frac{exp\left(m_v + \kappa_{k,v}^{(t)} + \kappa_{d,v}^{(c)} + \kappa_{y_d,k,v}^{(i)}\right)}{\sum_v^V exp\left(m_v + \kappa_{k,v}^{(t)} + \kappa_{d,v}^{(c)} + \kappa_{y_d,k,v}^{(i)}\right)}$$

where $\{\kappa_{\cdot,\cdot,\cdot}^{(t)} + \kappa_{\cdot,\cdot,\cdot}^{(c)} + \kappa_{\cdot,\cdot,\cdot}^{(i)}\}$ is a collection of topic (t), covariates (c) and interaction (i) coefficients. In this research, we included two covariates, the times of citation and the category of each article (see section 3.1 and Table 2) in our model. The times of citation were adjusted according to the year of publication to reduce the impact caused by the accumulation of citations over time.

$$Adjusted_{citations} = \frac{Total_{citations}}{(2019 - Publication_{year}) + 1}$$

We further specified that the times of citation and category were related to the topic prevalence and that the category alone was the content-related covariate.

The input text was lemmatised after removing stop words (besides the software package-provided list, some other terms were manually added, such as ‘this’, ‘this’, ‘etc.’ and ‘half’), format-related words in the abstract (purpose, design methodology, etc.), copyright text and punctuation. We noticed that some keywords containing important information are bigram terms, such as ‘big data’, ‘path analysis’, ‘clickstream data’, ‘e-commerce’ and ‘social media’. We concatenated those terms after lemmatisation. The candidate bigram terms were selected based on the keyword lists from the NAILS data set; additionally, we also selected some bigram terms with high frequencies in the text.

We utilised the held-out likelihood to select the number of topics, according to which 50% of the collected document contents are randomly selected to build the STM model, and the built models are evaluated based on their likelihood on the other 50% document contents that were not used to build the model. The same process was conducted with topic numbers from 2 to 20. Under each setting, 50 STM models were built, and the averaged held-out likelihood values were calculated. The averaged held-out likelihood from the STM models reached the optimal value when the topic number was equal to 14; thus, the number of topics was set to 14.

We first built 100 models with different initialisations and then selected the model with the highest average semantic coherence value. The average semantic coherence was a well-performing measure of topic model quality, which measured whether the top words in each topic tended to occur together across documents.

To quantify the impact of the found topics on citations, we performed an analysis of relative citation ratios, as follows. Given a population of documents arising out of several topics, the analysis assumed that citations arise based on the topical content and assigned credit (responsibility) for the citations of each document to the topics in proportion to their prevalence in the document; accordingly, $c_{d,k} = \theta_{d,k} \times c_d$ is the responsibility of topic k in a document d having topic proportions $\theta_{d,k}$ and citation count c_d. The average citation count arising from a particular topic is then the average of its assigned citations over the document population, $AC_k = \frac{1}{D}\sum_{d=1}^{D} c_{d,k}$. In contrast, the average prevalence of the topic over the document population is $\theta_k = \frac{1}{D}\sum_{d=1}^{D} \theta_{d,k}$. Finally, the relative citation ratio RC_k of a topic is simply its average citation count in the document collection divided by its average prevalence in the collection, $RC_k = AC_k/\theta_k$. The relative citation ratio indicates whether some topics are more likely to yield citations: if all topics gathered citations simply in proportion to how much they appear in the document collection, the relative citation ratios would be the same for all topics, whereas if the relative citation ratio is higher for some topics, it means the topic has yielded more citations than expected given its prevalence. We will report all three numbers θ_k, AC_k, and RC_k for each topic.

3.3.2 Comparison of STM versus LDA topic modelling

We chose the STM model over the more common LDA for two reasons: the better quantitative fit of STM to our document collection and the ability of STM to model covariates. We describe both below; we also briefly describe additional restrictions in the specific NAILS implementation of LDA.

Quantitative ability to model the document collection. As described in the previous subsection, held-out likelihood is a quantitative measure of how well a probabilistic model can fit a document collection, and we can use it to compare the STM topic model to the LDA model. LDA topic models can yield a held-out likelihood in the same way as STM models. Thus, we calculate the average held-out likelihood values of both the STM and LDA models over different numbers of topics. The STM models outperformed LDA models; the best held-out likelihood value from LDA models was -6.893, while the best value from STM models was -6.691.

In addition to the better quantitative criterion value, in practice, the STM results also seemed to be more semantically descriptive. Top words of the topics overlapped across topics (i.e. the topics did not have clearly separate characterisations); moreover, the top words of each topic were semantically relatively general compared to the topic contents found by STM.

Ability to model relationships between the text and covariates. LDA focuses on the text content alone, so it cannot provide insight into the relations between the text and other covariates, such as citations and category. Although conducting a post-analysis is possible, we chose to analyse the collected data (both text content and document-level covariates) in an integrated manner, which STM is directly able to do.

Additional restrictions on the NAILS software. Beyond the two general advantages of STM over LDA discussed above, the specific NAILS software implementation of LDA has additional restrictions that could be avoided by using another LDA software programme, which are briefly mentioned here: 1) NAILS only uses the text from the abstract for topic modelling and ignores other informative content of the data set, such as keyword columns; 2) NAILS uses ‘stemming’ (cutting suffixes and prefixes of words), whereas we preferred ‘lemmatisation’, which takes morphology into account and can find unified forms for more complicated cases, such as

irregular verbs (e.g. ‘drive’, ‘drove’ and ‘driven’ are stemmed differently from ‘drive’, ‘drove’ and driven’, but lemmatisation finds their common lemma ‘drive’). We stress that our reasons for choosing STM were the better quantitative performance and ability to model covariates, which hold regardless of the specific LDA implementation; the additional NAILS disadvantages are only mentioned for completeness.

The above advantages of STM over LDA motivated our choice of STM as the topic modelling method for the final analysis. We present the results of the analysis next.

4. Findings of the bibliometric analysis

4.1 Occurrence of the words

The frequency of lemmatised words (within keywords and the abstract) appearing in the 495 articles published between 2000 and 2018 is shown in Figure 1 for the top terms and in a rough overall graphical form as a word cloud in Figure 2. The frequencies show that the focus varies, for example, between the context of the studies (online) and the subject of the action (consumer, customer). In terms of context, the most studied environment seems to be online, including behaviour as well as searching for information regarding the products.

INSERT FIGURE 1 ABOUT HERE

INSERT FIGURE 2 ABOUT HERE

Figure 3 also shows the development of the occurrence of the keywords related to the analysis methods or the type of data discussed or used in the articles. Again, the lists of keywords as well

as abstracts of the articles were scrutinised. Based on this, the number of articles including clickstream, path analysis and text mining increased modestly from 2000 to 2018. Notably, however, the number of articles containing references to big data or discussing it experienced rapid growth between 2013 and 2018. In 2013, there was only one such article; in 2014, there were nine articles; in 2015, there were 20 articles; and in 2018, there were over 60 articles related to big data. Another growing area of research seems to be related to social media. Since 2013, there has been substantial growth in the number of articles, and in 2018 there were over 20 articles related to social media.

INSERT FIGURE 3 ABOUT HERE

When looking only at keywords, the popularity of big data was similarly evident. As Figure 4 shows, big data was listed as a keyword in 113 articles. In contrast, the more traditional objects of interest in this research field, such as e-commerce (26), clickstream data (16 articles) and online reviews (13), were much less in focus in the articles included in our study.

INSERT FIGURE 4 ABOUT HERE

In addition to the pure occurrence of the keywords, we also inspected their importance. To do this, we used the number of citations for the articles containing certain keywords. As can be seen in Figure 5, ‘Big data’ was the most-cited keyword, with 951 citations. ‘E-commerce’ (763 citations), ‘Internet’ (607) and ‘Clickstream data’ (544) were also widely cited keywords. Other keywords that had over 200 citations included ‘Internet marketing’ (310), ‘marketing’ (302), ‘Social media’ (287), ‘Consumer behavior’ (276) and ‘Computer-mediated environments’ (248).

INSERT FIGURE 5 ABOUT HERE

4.2 Development of the number of publications

Figure 6 shows the development of publications in terms of the number of articles over time. The number of articles published per year varied from one to 15 in the years between 2000 and 2010. After that, the growth was substantial, and the number of articles grew each year, from 10 published articles in 2011 to 128 in 2018.

INSERT FIGURE 6 ABOUT HERE

Also, looking at the relative publication volume (i.e. the fraction of publications included in our study of all publications within the Web of Science) reveals how the importance of the topics included in our study emerged. As Figure 7 shows, the fraction of all publications included in the study started to grow mildly from 2010 and turned into practically vertical growth from 2013 onwards.

INSERT FIGURE 7 ABOUT HERE

4.3 Noteworthy authors

Figure 8 provides the list of the most-cited authors. Of all the authors in the 495 articles, the six most cited are C. A. Lin , R. Larose, M. S. Eastin (all with 298 citations), S. J. Newell, B. A. Lafferty and R. E. Goldsmith (all with 285 citations). In addition to these six authors, four had over 200 citations: W. W. Moe, P. S. Fader, A. L. Montgomery and S. Bellman.

INSERT FIGURE 8 ABOUT HERE

In addition to focusing purely on the absolute number of citations, we also inspected the most important papers. They were identified utilising three importance measures: first, the In-Degree in the citation network; second, the citation count provided by the Web of Science (only for papers included in the data set); and finally, the PageRank score in the citation network. The top-ten highest-scoring papers were identified using these measures separately. The results were then combined, and duplicates were removed. The results presented in Table 3 are sorted by In-Degree, and ties are broken first by citation count and then by PageRank score.

INSERT TABLE 3 ABOUT HERE

Based on this analysis, the three most important papers in our data set were as follows: Montgomery et al. (2004), discussing the modelling of online browsing and path analysis using clickstream data; Bucklin et al. (2002), discussing research progress and future opportunities for modelling consumer choice on the Internet using clickstream data; and Johnson, Moe, Fader, Bellman and Lohse (2004), examining online search behaviour.

4.4 Important publication outlets

As Figure 9 shows, the most influential journals in terms of article count are the leading marketing journals, led by Marketing Science (14 articles) and followed by the Journal of Marketing (10). In addition, technology-oriented journals are potential outlets in this field: Electronics Commerce Research and Applications (10 articles), IEEE Access (8) and Expert Systems with Applications (10).

INSERT FIGURE 9 ABOUT HERE

The vital role of marketing journals is more apparent when the focus is on the volume of citations for different journals (see in Figure 10). Marketing Science is by far the most-cited

journal with 613 citations, followed by Journal of Marketing (469 citations) and Journal of Marketing Research (320). The next most-cited ones have been cited between around 300 times: Media Psychology (298 citations), Journal of Advertising (291), Management Science (265) and Journal of Business Research (260).

INSERT FIGURE 10 ABOUT HERE

5. Topic modelling

Next, we analysed the content of the text, including the titles, keywords and abstracts of the articles, to discover what kinds of topics might occur within our data set. Here, we conducted the text analysis by utilising a more advanced text analysis method called structural topic modelling, following the procedure described in section 3.3.

We finally extracted 14 meaningful topics and labelled them based on the top words of each topic, and the identified labels and corresponding top words are briefly described in the following. The topic *Shopping Experience and Satisfaction* contains top words including ‘consumer’, ‘study’, ‘intention’, ‘satisfaction’, ‘experience’ and ‘shop’. The topic *Online Behavior Pattern* includes the top words ‘online’, ‘use’, ‘time’ and ‘pattern’. The topic *Privacy in Digital Community* consists of words such as ‘information’, ‘social’, ‘datum’ and ‘privacy’. The topic *Information Search* comprises top words like ‘consumer’, ‘information’, ‘online’ and ‘search’. The topic *Challenges in Retail Market* contains words like ‘market’, ‘big data’, ‘research’, ‘retail’ and ‘challenge’. The topic *Quality of Customer Experience* puts more emphasis on words related to ‘customer’, ‘service’ and ‘quality’. The topic *Online as a Sales*

Channel includes top words such as ‘online’, ‘study’, ‘relationship’ and ‘sale’. The topic *Analysis of Social Networks* encompasses ‘datum’, ‘analysis’, ‘mobile’, ‘user’ and ‘network’. The topic *Brands and Online Reviews* consists of top words ‘product’, ‘use’, ‘review’ and ‘brand’. The topic *Purchasing Behavior* contains top words such as ‘purchase’, ‘behavior’ and ‘model’. The topic *Data Analytics and Food* includes the top words ‘datum’, ‘analytics’, ‘system’, ‘big data’, ‘business’, ‘value’, ‘process’, ‘new’ and ‘food’. The topic *Advertising Performance* contains top words such as ‘model’, ‘effect’, ‘use’ and ‘advertise’. The topic *Perceived Attitudes Towards Companies* is expressed by the top terms ‘model’, ‘attitude’ and ‘perceive’. The topic *Digital Age and Services* contains the terms ‘technology’, ‘service’, ‘trust’ and ‘knowledge’. The top word list for each topic and the corresponding word cloud figure can be found in Table 4 and Figure 11.

Out of the 14 identified topics, the topic *Privacy in Digital Community* has received the highest total number of citations (21.58). The topic *Shopping Experience and Satisfaction* has received the highest number of relative citations (2.93). The topic *Analysis of Social Networks* has received the fewest relative citations (2.06) compared to other topics. One possible explanation for this could be that this topic is related to research interests that have emerged only recently, and the small number of citations may be due to the nature of the accumulation of citations, although we have adjusted the citations over time.

The prevalence of each topic over each year is shown in Figure 12, where the red dotted line is the five-year moving average value. Based on the plot, although the specific prevalences vary year to year, there is some indication that the topics *Privacy in Digital Community*, *Quality of Customer Experience* and *Digital Age and Services* may similarly have a rising trend. Other topics are either unstable or stable over time, and no obvious trends are found. Note that there is

only one article in 2000; therefore, the drastic decreases or increases in topic prevalence during that year can only be due to the lack of data. We further conducted a Spearson non-parametric test between the topic prevalence and time on the above-mentioned three topics. The Spearman's rank correlation coefficient values are 0.58, 0.53 and 0.75, respectively, with corresponding p-values of 0.01, 0.02 and 0.0003. Another thing worth noticing is that the topic *Analysis of Social Networks* may have experienced a change point in the year 2006, with the trend of the topic prevalence decreasing until 2006 and increasing afterward. The Spearman's rank correlation coefficient value between time and topic prevalence of this topic from 2000 to 2006 is -0.85 (p-value: 0.02) and 0.84 (p-value: 0.0006) from 2006 to 2018.

We further examined the topic prevalence in different categories by specifying a linear regression model where the output was the proportion of the nine identified topics while the input was the category of the publication. The comparison of topic prevalences in different groups is displayed in Figure 13. As shown in the figure, the topic *Privacy in Digital Community* is more prevalent in the category 'Communication'. Fiore-Gartland and Neff's work (2015) focuses on 'data valence', which is relevant to the notion of this topic.

Meanwhile, the topic *Shopping Experience and Satisfaction* was frequently discussed in the literature under both the categories 'Hospitality' and 'Communication'. Studies in the above two categories have shared common interests regarding this topic. For example, Yuan et al.'s work (2008) from the category 'Hospitality' (which is categorised in the Web of Science as Hospitality, Leisure, Sport & Tourism) discussed customer attitudes and wine purchasing. The work of Bobkowski (2015), which belonged to the 'Communication' category, discussed how the personalities of news consumers affect their news-sharing behaviours.

For the topic *Online as a Sales Channel*, there was little obvious difference between the prevalences in different categories. One possible explanation for this could be that it is a general topic, so it received equal attention in different categories. This topic also contained a certain number of commonly used terms in this field, such as ‘study’, ‘sale’, ‘finding’ and ‘literature’.

The prominent prevalence of the topic *Digital Age and Services* in the categories ‘Health’ and ‘Environmental Sciences’ is also worth mentioning. It reveals the academic interest in digital services. Ramkumar et al. (2017) explored the potential of using mobile health data for orthopaedic surgeons, while Beal and Flynn’s work (2015) is an example of research on ‘smart metering’ in water utility, an important issue in environmental sciences.

The STM model also enabled us to compare the wordings of two different categories with respect to the same topic, in terms of which words or bigram terms were prevalent in the use of the topic within the different categories. We mainly compared the categories ‘Business and Economics’ and ‘Computer Science and Information Science’, as they had the largest number of publications (141 and 63, respectively, See Table 2), making the comparison meaningful.

The wording difference for the topic *Shopping Experience and Satisfaction* is displayed in Figure 14. The literature in the ‘Business and Economics’ category mentioned more terms related to consumers, such as ‘consumer’, ‘experience’ and ‘attitude’. In contrast, the terms mentioned in the ‘Computer Science and Information Science’ category seem to focus on the discussion of the overall trend and data sources, such as ‘positive’, ‘negative’, ‘sns’ and ‘news’.

There was also a difference between the categories ‘Business and Economics’ and ‘Computer Science and Information Science’ with respect to the topic *Privacy in Digital Community* (shown in Figure 15). Researchers in ‘Computer Science and Information Science’ tended to place more emphasis on technical discussions, thus using terms such as ‘Communication’, ‘datum’ and ‘digital’ frequently. Meanwhile, terms related to human–computer interaction, such as ‘information’, ‘interaction’ and ‘social’, were found in the literature in the ‘Business and Economics’ category.

INSERT TABLE 4 ABOUT HERE

INSERT FIGURE 11 ABOUT HERE

INSERT FIGURE 12 ABOUT HERE

INSERT FIGURE 13 ABOUT HERE

INSERT FIGURE 14 ABOUT HERE

INSERT FIGURE 15 ABOUT HERE

6. Discussion

In this study, we report the evolution of scientific research with regard to the usage of large data sets on online consumer behaviour between 2000 and 2018 in terms of publications available in the ISI Web of Science database. The results based on 495 articles provide an overview of the existing information regarding research on online consumer behaviour utilising or discussing large data sets. This study provides information on the evolution of the field by identifying key

publications and authors and examining how certain topics have evolved over time. By utilising topic modelling and text analytic techniques, we also identified certain research themes from the papers included in our data. This analysis offers a guide to those who want to contribute to this field of online consumer research by providing information, for example, about which journals should be consulted, which authors are the most prominent and what topics are relevant in different fields.

With this paper, we also contribute to the methodology related to literature surveys and bibliometric analyses. Notably, we utilise a more advanced methodology beyond traditional bibliometric analyses—or even beyond basic LDA modelling. We conducted topic modelling to extract the latent topics from the collected literature. We demonstrated the usage of STM, a more advanced text analysis tool in literature review studies, showing that this automated framework enables researchers to gain more interesting insights than with other off-the-rack text-mining tools such as LDA.

Our results show that the research landscape of online consumer behaviour utilising large data sets has evolved during the selected time period (2000 to 2018), so the topic has evidently become more popular and important in recent years. The number of articles published per year has increased rapidly from 2011 onwards, and the share of the articles from all publications within the Web of Science has skyrocketed since 2013. Perhaps the most notable change relates to the occurrence of the term ‘big data’ in the papers. This is particularly obvious from 2013 onwards, and it appears to be a highly dominant area of research. Moreover, the number of citations related to ‘big data’ is dominant, even though researchers have only relatively recently started to focus on big data. In addition, articles focusing on the Internet in general (Internet, e-commerce) and clickstream data as well as on Internet marketing and consumer behaviour are

among the most-cited articles. However, the popularity of big data is so evident that we consider it likely the number of citations for such publications will continue to increase considerably.

When we look at the publication outlets, it seems that the leading marketing journals are those that conscious academics should target in order to exert as much influence within the research community as possible. Our results show that journals such as Marketing Science and the Journal of Marketing and the Journal of Marketing Research have published most of the articles, and they are also among the journals that are cited most often. However, general management and business journals (e.g. Management Science and Journal of Business Research) as well as journals with a technological orientation (e.g. IEEE Access and Expert Systems with Applications) are other potential publication outlets for the researchers in this field.

In terms of topic modelling, by utilising the above-mentioned novel and more advanced STM methodology, we managed to identify certain topics under which the papers included in the data set could be classified. Using this kind of modelling, it is possible to show hidden common content in papers by studying latent topics. It is also possible to compare citations of different topics or the occurrence of the topics in different fields of research. As a result, we discovered that papers related to *Privacy in Digital Community* had received the highest number of citations, followed by papers under the topics *Challenges in Retail Markets* and *Shopping Experience and Satisfaction*. Thus, it seems that these are the most influential topics in the field. A bit surprisingly, papers under the topic *Analysis of Social Networks* have received fewer citations. Again, however, we believe that due to the popularity of articles related to data analytics and social media, together with the nature of the accumulation of citations, this topic could receive a notable number of citations in the future (indeed, since 2006, the trend of the topic prevalence has been increasing). One interesting finding related to the topic *Shopping Experience and*

Satisfaction was that in the field of 'Business and Economics', the papers seemed to focus on consumer perspective (customer, experience etc.), whereas papers in the field of 'Computer Science and Information Sciences' were more related to the technical side of the phenomenon (e.g. utility, positive, negative).

Interestingly, the rise of research related to big data seems to be quite general in nature and not heavily focused on any specific topic. This conclusion can be made based on our analysis of the topic prevalence over time, which showed that even though there are increasing trends in some of the topics ('*Privacy in Digital Community*', '*Quality of Customer Experience*' and '*Digital Age and Services*'), the increase is not as substantial as in the number of articles related to big data, which experienced drastic growth between 2013 and 2018.

6.1 Directions for the future research in online consumer behaviour

As stated above and as verified by the results of this paper, larger data sets, especially in the form of big data, are being used more commonly, and there is not a problem with regard to getting the data. However, big data is worthless on its own, and the potential value from it is gained only when it is leveraged to drive decision-making (Gandomi & Haider, 2015). Among other issues related to big data, the Marketing Science Institute (2016) listed certain research priorities for the years 2016–2018, including the integration of behavioural and marketing theories with big data as well as utilising multiple sources and types of information to gain insights for decision-making. Thus, the most likely problem for both scholars and practitioners is how to utilise big data more efficiently, and the question is how to make sense of what is out there (Van Auken, 2015). As Gandomi and Haider (2015) argued, although major innovations in analytical

techniques for big data have not yet occurred, the emergence of such novel analytics (e.g. real-time analytics) is likely in the near future. The focus in research as well as in analytics within companies should also shift from *what* consumers do to *why* they do something, that is, to understanding their behaviour. To do this effectively, multiple sources of data need to be incorporated (see Shlomo & Goldstein, 2015), and scholars are urged to determine whether to rely only on data collected online or whether some kind of offline contextual data should be used as complementary data (Mahrt & Scharkow, 2013; see also Murty, 2008 and Orgad, 2009). The use of several sources of data would also help to overcome the problems of drawing conclusions based on aggregate-level big data. For instance, statistical relationships on the aggregate level cannot straightforwardly be applied at the individual level, as there is a risk of an ecological fallacy (see e.g. Brewer & Venaik, 2014), that is, the error of assuming that observations in aggregate data also occur on the individual level (Mahrt & Scharkow, 2013).

6.2 Limitations and research opportunities

The main limitation in this paper is that the data were collected only from the ISI Web of Science database. For further research, scholars might want to consider conducting a bibliometric analysis using other databases as well, such as Scopus, which also contain non-indexed journals that are unavailable in the ISI Web of Science. More articles could also be found on Google Scholar, for example, which includes not only citations in journals but also citations in other academic papers available on the Internet. Related to this, future research could also compare the results obtained from other databases against the results of this study.

References

Beal, C. D., & Flynn, J. (2015). Toward the digital water age: Survey and case studies of Australian water utility smart-metering programs. *Utilities Policy*, *32*, 29–37.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77. http://doi.org/10.1145/2133806.2133826

Brewer, P., & Venaik, S. (2014). The ecological fallacy in national culture research. *Organization Studies*, *35*(7), 1063–1086.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Seventh International World-Wide Web Conference (WWW 1998)*, April 14–18, 1998, Brisbane, Australia.

Bobkowski, P. S. (2015). Sharing the news: Effects of informational utility and opinion leadership on online news sharing. *Journalism & Mass Communication Quarterly*, *92*(2), 320–345.

Bucklin, R. E., Lattin, J. M., Ansari, A., Gupta, S. Bell, D., Coupey, . . . Steckel, J. (2002). Choice and the internet: From clickstream to research stream. *Marketing Letters*, *13*(3), 245–258.

Bucklin, R. E., & Sismeiro, C. (2009). Click here for internet insight: Advances in clickstream data analysis in marketing. *Journal of Interactive Marketing*, *23*, 35–48.

Cantallops, A. S., & Salvi, F. (2014). New consumer behavior: A review of research on eWOM and hotels. *International Journal of Hospitality Management*, *36*, 41–51.

Chatterjee, P., Hoffman, D. L., & Novak, T. P. (2003). Modeling the clickstream: Implications for web-based advertising efforts. *Marketing Science*, *22*(4), 520–541.

Chen, H., Chiang, R., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*(4), 1165–1188.

Chintagunta, P. K. (2001). Endogeneity and heterogeneity in a probit demand model: Estimation using aggregate data. *Marketing Science*, *20*(4), 442–456.

Corrigan, H. B., Craciun, G., & Powell, A. M. (2014). How does Target know so much about its customers? Utilizing customer analytics to make marketing decisions. *Marketing Education Review*, *24*(2), 159–166.

Darley, W. K., Blankson, C., & Luethge, D. J. (2010). Toward an integrated framework for online consumer behavior and decision making process: A review. *Psychology & Marketing*, *27*(2), 94–116.

De Mauro, A., Greco, M., Grimaldi, M., & Ritala, P. (2018). Human resources for big data professions: A systematic classification of job roles and required skill sets. *Information Processing and Management, 54*(5), 807-817.

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big data consumer analytics and the transformation of marketing. *Journal of Business Research*, *69*, 897–904.

Fetscherin, M., & Heinrich, D. (2015). Consumer brand relationship research: A bibliometric citation meta-analysis. *Journal of Business Research*, *68*, 380–390.

Fetscherin, M., & Usunier, J. C. (2012). Corporate branding: An interdisciplinary literature review. *European Journal of Marketing*, *46*(5), 6–44.

Fiore-Gartland, B., & Neff, G. (2015). Communication, mediation, and the expectations of data: Data valences across health and wellness communities. *International Journal of Communication*, *9*, 19.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35, 137–144.

Goh, K. Y., Chu, J., & Wu, J. (2015). Mobile advertising: An empirical study of temporal and spatial differences in search behavior and advertising response. *Journal of Interactive Marketing*, *30*, 34–45.

Ha, Y., Kwon, W. S., & Lennon, S. J. (2007). Online visual merchandising (VMD) of apparel web sites. *Journal of Fashion Marketing Management*, *11*(4), 477–493.

Hajikhani, A. (2017). Emergence and dissemination of ecosystem concept in innovation studies: A systematic literature review study. *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Hauser, J. R., Urban, G. L., Liberali, G., & Braun, M. (2009). Website morphing. *Marketing Science*, *28*(2), 202–223.

Hausman, A. V., & Skiepe, J. S. (2009). The effect of web interface features on consumer online purchase intentions. *Journal of Business Research*, *62*(1), 5–13.

Hofacker, C. F., Malthouse, E. C., & Sultan, F. (2016). Big data and consumer behavior: Imminent opportunities. *Journal of Consumer Marketing*, *33*(2), 89–97.

Huang, P., Lurie, N. H., & Mitra, S. (2009). Searching for experience on the web: An empirical examination of consumer behavior for search and experience goods. *Journal of Marketing*, *73*, 55–69.

Johnson, E. J., Moe, W. W., Fader, P. S., Bellman, S., & Lohse, G. L. (2004). On the depth and dynamics of online search behavior. *Management Science, 50*(3), 299–308.

Kim, M., & Lennon, S. (2008). The effects of visual and verbal information on attitudes and purchase intentions in internet shopping. *Psychology & Marketing*, *25*, 146–178.

Knutas, A., Hajikhani, A., Salminen, J., & Porras, J. (2015). Cloud-based bibliometric analysis service for systematic mapping studies. *CompSysTech 2015*.

Koufaris, M. (2002). Applying the technology acceptance model and flow theory to online consumer behavior. *Information Systems Research*, *13*(2), 205–223.

Lee, S., & Chen, L. (2010). The impact of flow on online consumer behavior. *The Journal of Computer Information Systems*, *50*(4), 1–10.

Li, H., & Kannan, P. K. (2014). Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. *Journal of Marketing Research*, *51*(1), 40–56.

Lycett, M. (2013). 'Datafication': Making sense of (big) data in a complex world. *European Journal of Information Systems*, *22*(4), 381–386.

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, *57*(1), 20–33.

Manganari, E. E., Siomkos, G. J., Rigopoulou, I. D., & Vrechopoulous, A. P. (2011). Virtual store layout effects on consumer behaviour: Applying an environmental psychology approach in the online travel industry. *Internet Research*, *21*(3), 326–346.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Marketing Science Institute (2016). *Research priorities 2016–2018*. Cambridge, MA: Marketing Science Institute.

Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*, *13*(1/2), 29–39.

Moe, W. W. (2006). An empirical two-stage choice model with varying decision rules applied to internet clickstream data. *Journal of Marketing Research*, *43*(4), 680–692.

Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data, *23*(4), 579–595.

Mummalaneni, V. (2005). An empirical investigation of web site characteristics, consumer emotional states and online shopping behaviors. *Journal of Business Research*, *58*(4), 526–532.

Murty, D. (2008). Digital ethnography. An examination of the use of new technologies for social research. *Sociology*, *42*(5), 837–855.

Nicolaisen, J. (2010). Bibliometrics and citation analysis: From the Science Citation Index to cybermetrics. *Journal of the American Society for Information Science and Technology*, *61*(1), 205–207.

Orgad, S. (2009). How can researchers make sense of the issues involved in collecting and interpreting online and offline data? In A. N. Markham & N. K. Baym (Eds.), *Internet inquiry*. *Conversations about method* (pp. 33–53). Los Angeles, CA: Sage.

Philander, K., & Zhong, Y. (2016). Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. *International Journal of Hospitality Management*, *55*(2016), 16–24.

Press, P. (2013). A very short history of big data. Retrieved from http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

Ramkumar, P. N., Muschler, G. F., Spindler, K. P., Harris, J. D., McCulloch, P. C., & Mont, M. A. (2017). Open mHealth architecture: A primer for tomorrow's orthopedic surgeon and introduction to its use in lower extremity arthroplasty. *The Journal of Arthroplasty*, *32*(4), 1058–1062.

Richard, M.-O., & Habibi, M. R. (2016). Advanced modeling of online consumer behavior: The moderating roles of hedonism and culture. *Journal of Business Research*, *69*, 1103–1119.

Roberts, M. E. (2014). Fear or friction? How censorship slows the spread of information in the digital age. *Unpublished manuscript*, *26*.

Ruh, B. (2012). The industrial internet: Even bigger than big data. Retrieved from http://www.forbes.com/sites/ciocentral/2012/10/04/the-industrial-internet-even-bigger-than-big-data/

Senecal, S., Kalczynski, P. J., & Nantel, J. (2005). Consumers' decision-making process and their online shopping behavior: A clickstream analysis. *Journal of Business Research, 58*(11), 1599–1608.

Shlomo, N., & Goldstein, H. (2015). Editorial: Big data in social research. *Journal of the Royal Statistical Society*, *178*(4), 787–790.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces: 63–70.* Retrieved from http://www.aclweb.org/anthology/W/W14/W14-3110

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research, 70,* 263–286.

Torres, E. N., & Singh, D. (2016). Towards a model of electronic word-of-mouth and its impact on the hotel industry. *International Journal of Hospitality & Tourism Administration*, *17*(4), 472–489.

Tvinnereim, E., & Fløttum, K. (2015). Explaining topic prevalence in answers to open-ended survey questions about climate change. *Nature Climate Change*, *5*(8), 744–747.

Van Auken, S. (2015). From consumer panels to big data: An overview on marketing data development. *Journal of Marketing Analytics, 3*(1), 38–45.

Van den Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research*, *166*(2), 557–575.

Yuan, J., Morrison, A. M., Cai, L. A., & Linton, S. (2008). A model of wine tourist behaviour: A festival approach. *International Journal of Tourism Research*, *10*(3), 207-219.

Zhai, C. (2017). Probabilistic topic models for text data retrieval and analysis. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1399–1401.

Table 1
Search Criteria

Level	Search term
Data	TOPIC: (“big data” OR “clickstream” OR “text data” OR “path analysis” OR “customer analytics”)
	AND
Individual	TOPIC: (customer OR consumer)
	AND
Related	TOPIC: (“purchase” OR “buying” OR “online discussion” OR “information search” OR "customer journey” OR “online” OR “mobile”)

Table 2
Number of Publications in Each Category

Category	Number of publications
Business & Economics	195
Computer Science and Information Science	133
Engineering	33
Health	26
Hospitality	26
Environmental Sciences	16
Communication	14
Other	35

Table 3
Most Important Papers

Year	Author	Title	Journal	InDegree	Times Cited	PageRank
2004	Montgomery et al.	Modeling online browsing and path analysis using clickstream data	MARKETING SCIENCE	48	192	0.0000810
2002	Bucklin et al.	Choice and the Internet: From clickstream to research stream	MARKETING LETTERS	23	55	0.0000626
2004	Johnson et al.	On the depth and dynamics of online search behavior	MANAGEMENT SCIENCE	22	206	0.0000621
2003	Chatterjee et al.	Modeling the clickstream: Implications for Web-based advertising efforts	MARKETING SCIENCE	20	167	0.0000574
2005	Van den Poel et al.	Predicting online-purchasing behaviour	EUROPEAN JOURNAL OF OPERATIONAL RESEARCH	13	89	0.0000538
2009	Bucklin et al.	Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing	JOURNAL OF INTERACTIVE MARKETING	13	76	0.0000544
2005	Senecal et al.	Consumers decision-making process and their online shopping behavior: a clickstream analysis	JOURNAL OF BUSINESS RESEARCH	10	71	0.0000508
2014	Li et al.	Attributing Conversions in a Multichannel Online Marketing Environment: An Empirical Model and a Field Experiment	JOURNAL OF MARKETING RESEARCH	10	62	0.0000514
2006	Moe et al.	An empirical two-stage choice model with varying decision rules applied to Internet clickstream data	JOURNAL OF MARKETING RESEARCH	10	47	0.0000495
2009	Huang et al.	Searching for Experience on the Web: An Empirical Examination of Consumer Behavior for Search and Experience Goods	JOURNAL OF MARKETING	9	201	0.0000498

Table 4
Extracted Topics from STM

Topic **(Prevalence (%)/ Average Citations/Relative Citation*)**	**Top words**
Shopping Experience and Satisfaction (6.86/20.12/2.93)	consumer, study, intention, satisfaction, experience, shop, examine, usage, positive, understand
Online Behavior Pattern (7.68/17.89/2.33)	online, use, time, pattern, base, propose, choice, good, framework, across
Privacy in Digital Community (7.91/21.58/2.73)	information, social, datum, privacy, digital, design, content, internet, socialmedia, change
Information Search (7.91/20.40/2.58)	consumer, information, online, search, price, user, behavior, internet, use, website
Challenges in Retail Markets (8.19/21.32/2.60)	market, bigdata, research, retail, challenge, store, author, management, source, potential
Quality of Customer Experience (7.72/16.50/2.14)	customer, service, quality, product, mine, approach, buy, strategy, algorithm, loyalty
Online as a Sales Channel (6.50/16.30/2.51)	online, study, relationship, sale, behavior, channel, finding, demand, role, literature
Analysis of Social Networks (6.91/14.23/2.06)	datum, analysis, mobile, user, network, method, customer, paper, sentiment, text
Brands and Online Reviews (7.19/18.79/2.61)	product, use, review, brand, impact, predict, show, research, result, variable
Purchasing Behavior (6.80/15.03/2.21)	purchase, behavior, model, customer, website, decision, visit, ecommerce, market, provide
Data Analytics and Food (7.16/16.94/2.37)	datum, analytics, system, bigdata, business, value, process, new, food, approach
Advertising Performance (6.61/18.80/2.84)	model, effect, use, advertise, performance, result, firm, increase, purchase, web
Perceived Attitudes Towards Companies (6.16/14.89/2.42)	model, attitude, perceive, company, industry, test, mobile, toward, retailer, socialmedia
Digital Age and Services (6.42/15.08/2.35)	technology, service, trust, knowledge, health, study, paper, collect, app, factor

*Relative Citation = Average Citations/Prevalence

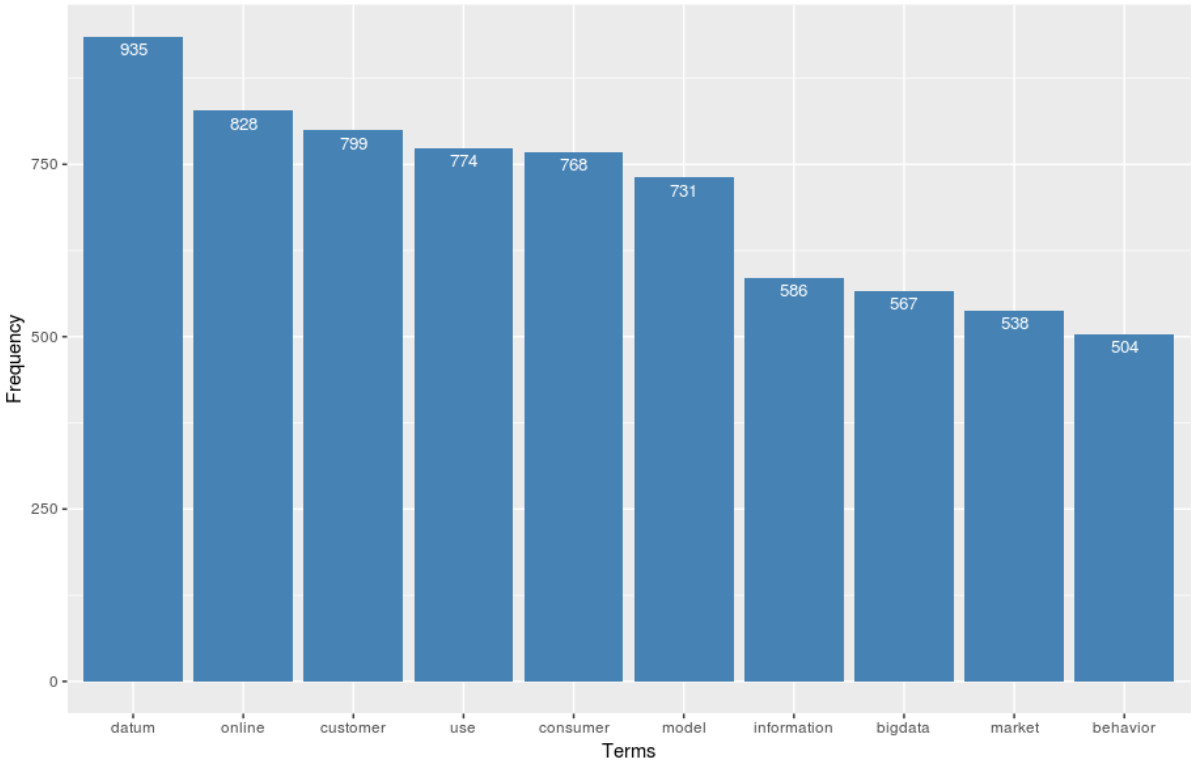

Figure 1. Top Terms

Figure 2. Word Cloud of All Literatures

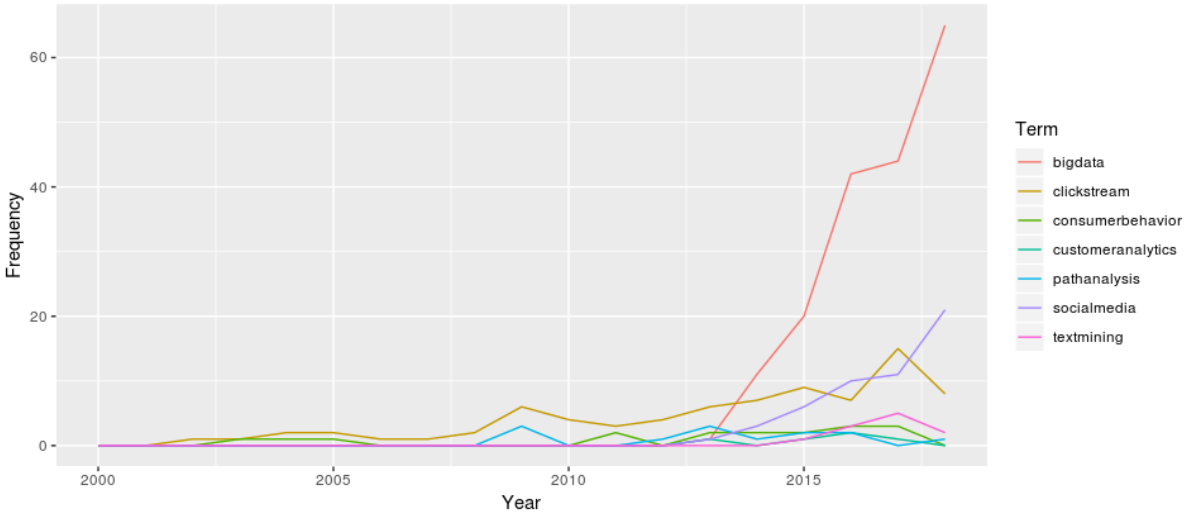

Figure 3. Term Frequency Over Time

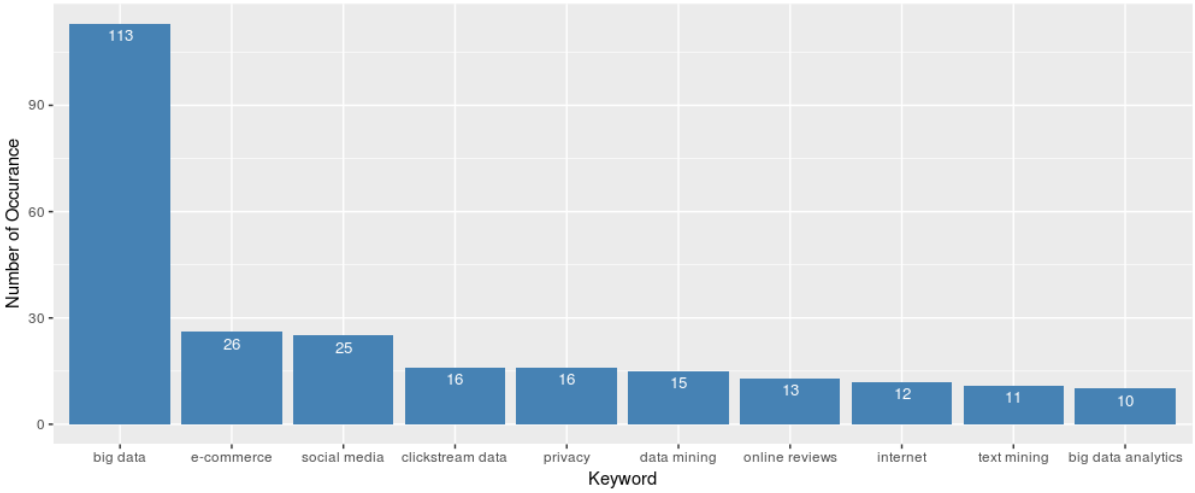

Figure 4. Keyword Occurence

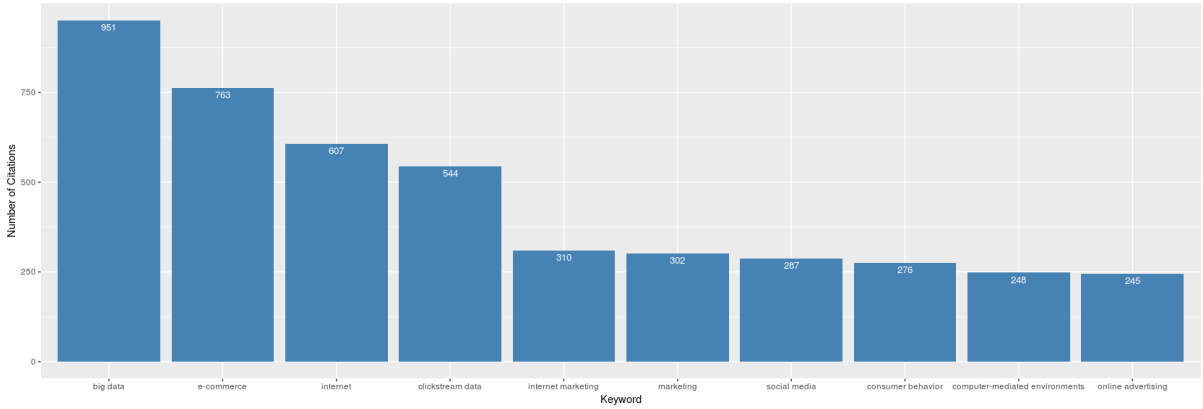

Figure 5. Keyword Citations

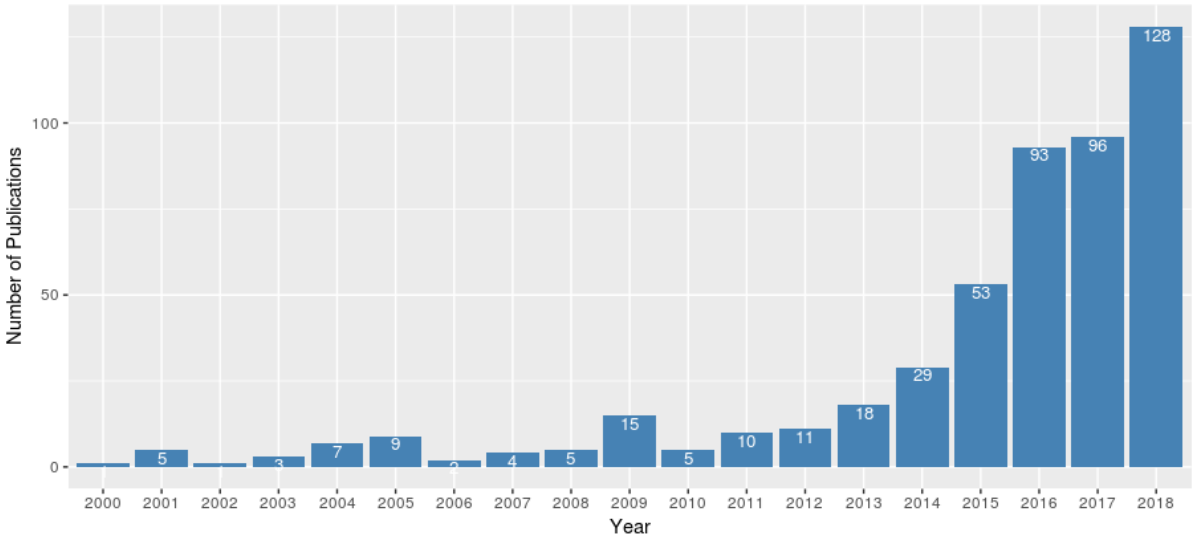

Figure 6. Number of Publications Between 2000 and 2018

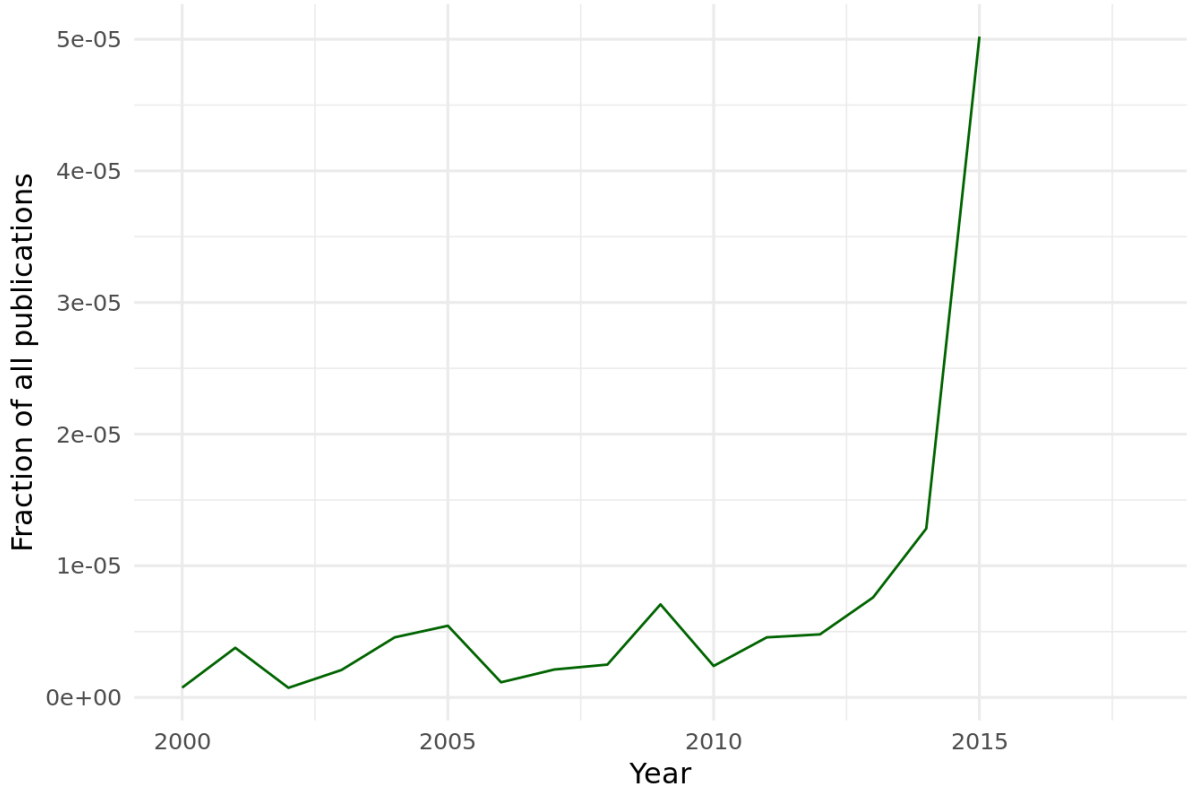

Figure 7. Relative Publication Volume

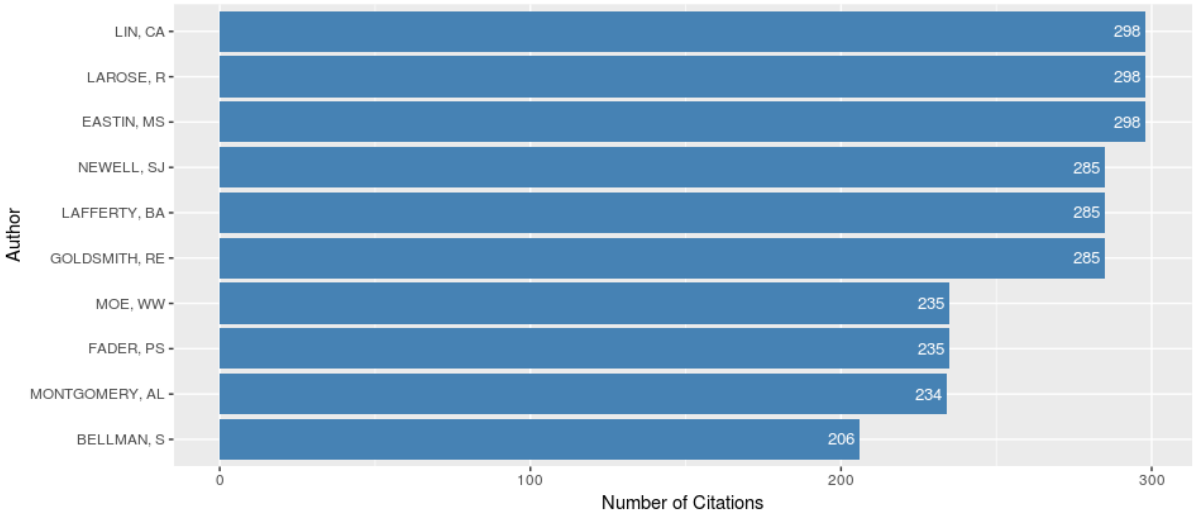

Figure 8. Top 10 Cited Authors

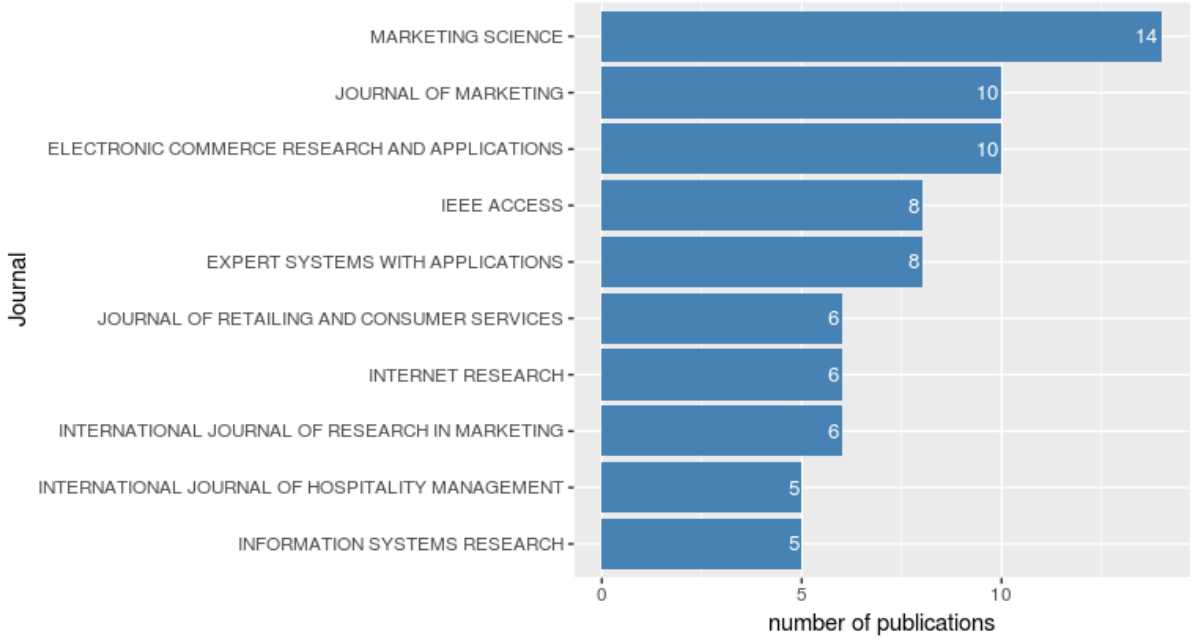

Figure 9. Top 10 Popular Journals

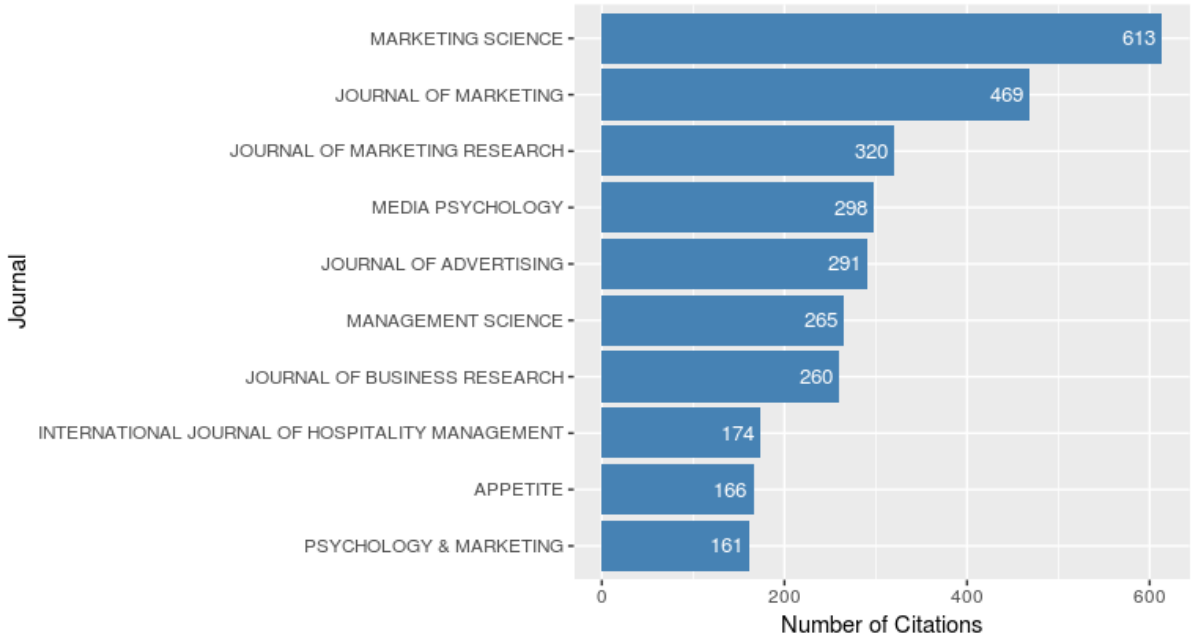

Figure 10. Top 10 Cited Journals

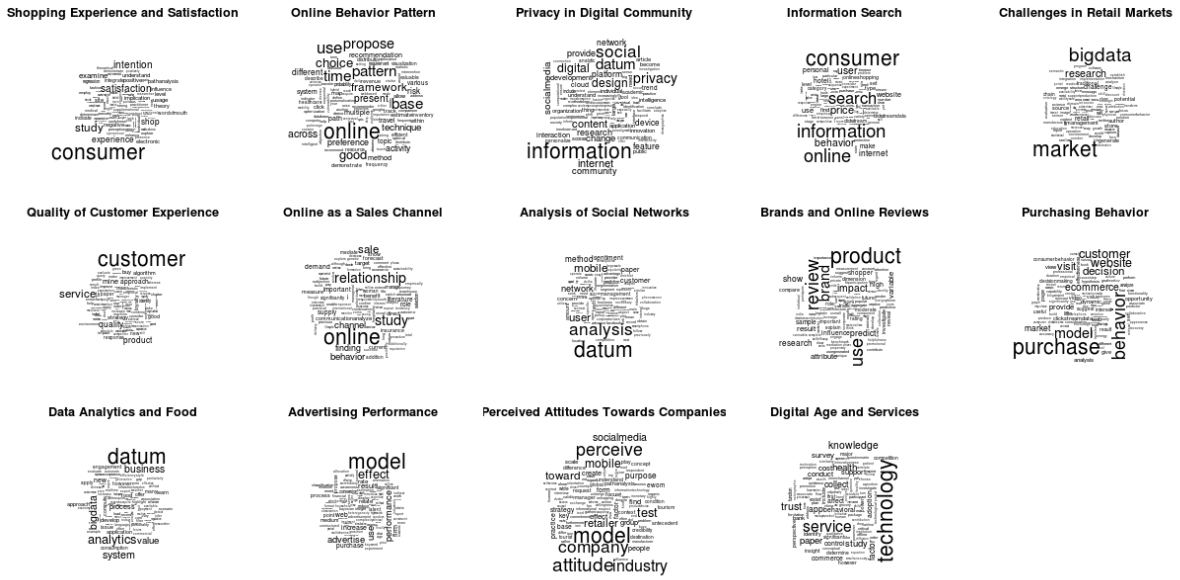

Figure 11. Word Cloud of each Topic from the STM Model

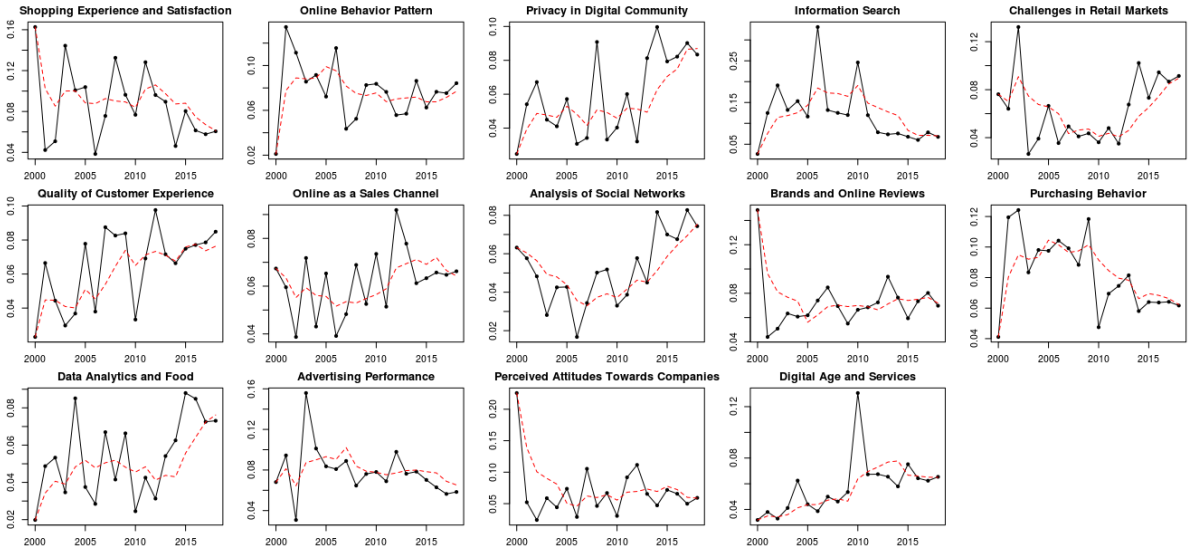

Figure 12. Topic Prevalence Over Time

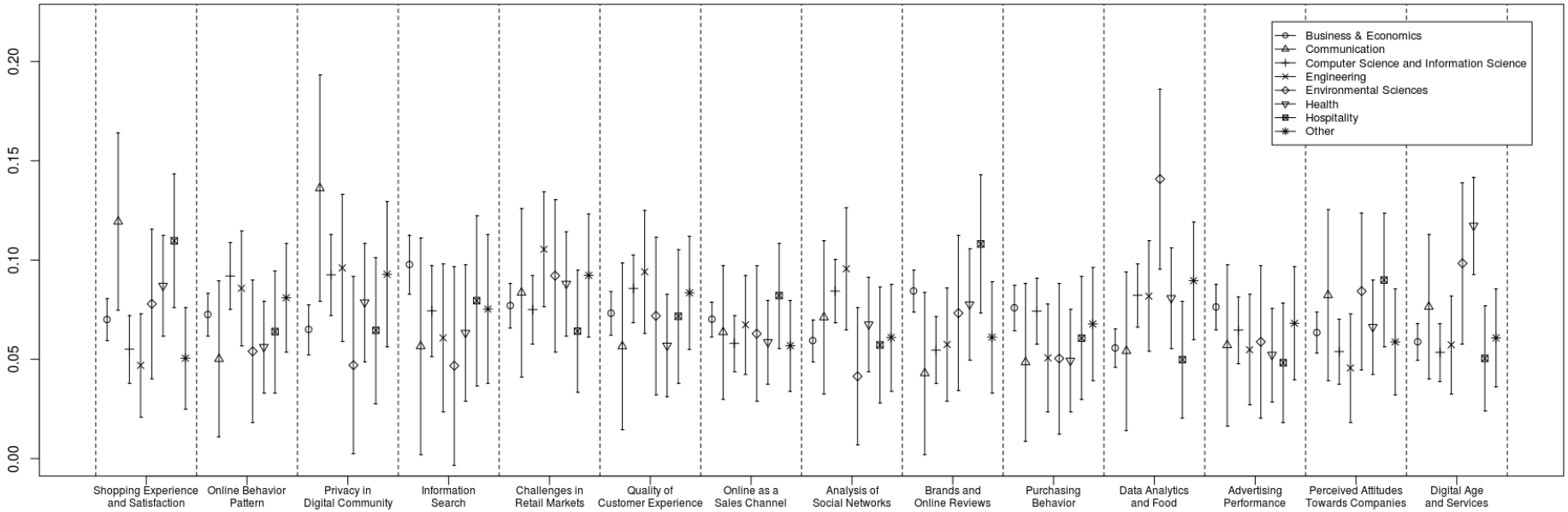

Figure 13. Topic Prevalence Over Categories

Figure 14. Wording difference in Topic *Shopping Experience and Satisfaction*

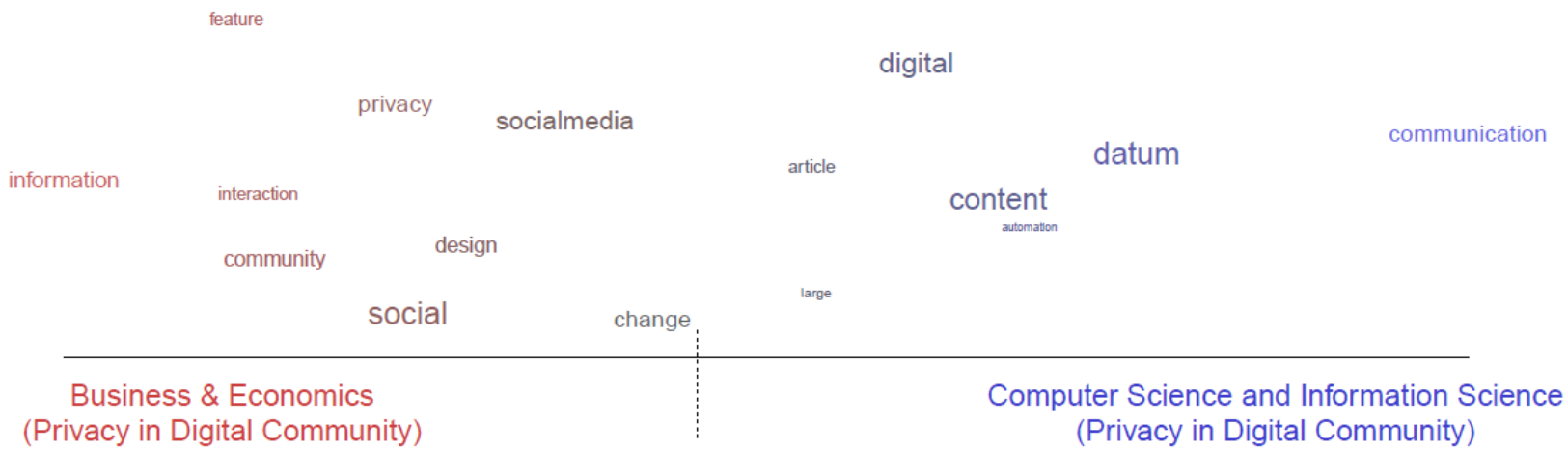

Figure 15. Wording difference in Topic *Privacy in Digital Community*