Marketing Models for Customer Engagement Behaviors by Using Large Scale and Unstructured Data

著者	IGARASHI MIRAI
学位授与機関	Tohoku University
学位授与番号	11301甲第19605号
URL	http://hdl.handle.net/10097/00131012

TOHOKU UNIVERSITY

DOCTORAL THESIS

Marketing Models for Customer Engagement Behaviors by Using Large Scale and Unstructured Data

Author:
Mirai IGARASHI

Supervisor:
Dr. Nobuhiko TERUI

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Graduate School of Economics and Management

January 4, 2021

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor, Prof. Nobuhiko Terui, who enthusiastically supported me a lot since undergraduate school. He was always helpful in discussing my research and his precise and insightful suggestions pushed me to sharpen my thinking and brought my work to a higher level. Without his valuable guidance and persistent help, this dissertation would not have been possible.

I would also like to thank two members of my dissertation committee: P. K. Kannan and Tsukasa Ishigaki. Prof. Kannan gave grateful guidance during my visit at the University of Maryland, and he always encouraged me with many positive words. Prof. Ishigaki not only taught me how to conduct the research from the methodology perspective but also gave many valuable pieces of advice even on non-research-related topics such as writing good applications to academic grants and constructing the network of Ph.D. students.

Furthermore, I would like to thank Prof. Kunpeng Zhang of the University of Maryland, Prof. Yasumasa Matsuda, Prof. Yoshimasa Uematsu, and Prof. Yinxing Li of the graduate school of economics and management in Tohoku University, and Dr. Toshikuni Sato and Dr. Aijing Xing of graduates of the department, for all valuable guidance, advice, and discussions.

Last but not least, I would like to thank my family, all of my friends, and especially Hikari who always encouraged me to go on.

Preface

Modern consumers often use social media and e-commerce platforms to express their opinions on products and services and to deepen their relationships with companies and brands. Researchers call these behaviors *customer engagement behaviors* (CEB), which have been attracting much attention recently in various fields of marketing, consumer behavior, and sociology because of their different backgrounds and nature compared with the traditional data handled in these fields, such as the recorded purchase behaviors and questionnaire responses. Furthermore, the CEB data contain a wealth of information such as reasons for not making a purchase and post-purchase impressions. The cost of data collection is low, thanks to the development of information processing technology, thereby increasing their marketing value.

In the fields of statistics and information science, researchers have established analysis methods to the CEB data, such as the deep learning models and topic modeling approaches. However, in the field of social science, especially marketing, such analysis methods have not yet been established. In the social media era, there is a growing need to make use of the CEB data to better understand consumer behaviors and create effective marketing activities. Thus, it is important to establish the analysis methods to the CEB data. Why are there no established methods for analyzing the CEB data? The main reason is that most CEB data consist of unstructured information such as text and images, which are difficult to quantify. Moreover, the scale is often very large because many consumers behave in a variety of ways and are associated with each other at the same time.

The machine learning approaches used in the statistical field can deal with such unstructured and large-scale data, but they are based on the idea that the purpose of the analysis is to predict and summarize. Further, the model structure can be black box as long as it produces good results. Some econometric models used in the marketing field are aimed at understanding the driving factors and spillover effects of such consumer behaviors; hence, the model structure must be sophisticated to ensure that the findings from the analysis help to achieve such purposes. However, the CEB data consisting of unstructured and large-scale data cannot be handled by the econometric models because they assume to deal with structured and relatively small data such as sales data. Therefore, I address the above unsolved and important issues in the marketing field through the development of new marketing models for the CEB data analysis. I developed the models by applying and reconstructing machine learning methods while retaining effective model structures to understand the driving factors and the spillover effects of consumer behaviors.

In Chapter 1, I introduce a network model considering the text information on social media to simultaneously understand the community structure on the network and the communities' topics of interest by the social media users. Identifying the community and topic structure in social media data helps us understand why people relate to other friends and post contents on the media, that is, the driving factors for the CEB on social media. Through model comparison, this study also clarifies the effects of considering the text and the network information on the performance for the community structure recovery. Moreover, it shows that the proposed model can find realistic and meaningful community structures from large online networks through an empirical analysis using the Twitter dataset.

In Chapter 2, I extend the network model introduced in Chapter 1 by differentiating the edge generation probability for each node to consider the node degree heterogeneity, which is often observed in a real social network (e.g., influential users). The empirical analysis using Twitter dataset shows that the model can provide interpretable community and topic structure from the network data. Moreover, it discusses the effects of the simultaneous consideration of the network and text information on the estimation results and the predictive performance. The discussion is more detailed than that in Chapter 1 because of several model comparisons with the independent approach dealing with data separately and comparative models subtracting the model features considering text information and degree heterogeneity.

In Chapter 3, I examine the spillover effects, or social influence, of the content-generating behaviors on social media, unlike the above two studies addressing the driving factors. This study contributes to a large body of the literature on social influence by identifying the differences in social influence across topics of the user-generated contents on social media and simultaneously estimating the dimensions of the topics and social influences varying for the topics using the proposed dynamic topic model. In an empirical analysis using the Pinterest data, the proposed model extracts interpretable topics from the image data, captures heterogeneity in topic proportions considering the time evolution, and estimates different social influences for the topics between the same pair of users.

The above three studies discuss the CEB on social media, but another important CEB, writing customer reviews, is also observed. In Chapter 4, I introduce an extended topic model combining the preliminary expertise in the product domain into the topic assignment model to address the difficulty in identifying product attributes in the review text using the conventional topic model. The empirical study using Amazon dataset shows that the proposed model, with statistical limitations, can improve the interpretability of the identified product attributes while showing comparable generalization performance to the unrestricted model. Furthermore, the model provides some interesting findings about the relationships between the product attributes in the review and product satisfaction and review helpfulness. For example, the "ingredient" topic in reviews decreases the level of satisfaction and perceived helpfulness, whereas the "health" topic increases the levels of both.

In Chapter 5, examining on customer review analysis, I introduce a combined model of word embedding and topic modeling to address the major issue of conventional topic modeling that ignores the word order and hence does not consider the context of the text information. The combination of the word embedding and topic modeling itself has been proposed in the literature. However, I extend this approach from two perspectives: (i) the supervised learning to understand the effects of the topic structure in the review text on the review ratings and (ii) the consideration of the text sentiment to determine the product attributes in the text by considering the sentiment proportion obtained through the sentiment analysis on the review text. Moreover, the proposed model sophisticates the model structure for the preference measurement by assuming brand heterogeneity in the effects of the attribute proportion and by considering the direct and indirect effects of consumer attributes on the overall satisfaction. The empirical study using the Sephora dataset shows that the proposed model outperforms the generalization error comparison with comparative models. Moreover, it provides some interpretable attributes (e.g., flaking, smell, and

eyelash performance of the mascara product) and interesting findings on the preference structure, such as the heterogeneous impacts of the attribute proportion on brand satisfaction and the positive and direct effect of receiving free samples on the satisfaction.

Contents

List of Figures

List of Tables

Chapter 1

Characterization of Topic-based Online Communities by Combining Network Data and User Generated Content

1.1 Introduction

The product or information diffusion is affected by not only the communication between companies and consumers but also by interactions between consumers such as word-of-mouth on social media or product reviews on e-commerce sites; the impact of the latter is stronger in the modern social media development. Companies are required to implement various marketing activities considering such relationships between customers. A significant first step towards learning about the relationship between customers is to grasp their community structure on networks. If nodes of a network can be divided into some (potentially overlapping) groups such that nodes are densely connected internally, the network is said to have a community structure. Furthermore, researchers know that some network structures with closely connected nodes, or customers, can bring some benefits to companies such as sharing contents (Peng et al., 2018), achieving long-term popularity (Ansari et al., 2018), and accelerating product innovation (Peres, 2014). Therefore, uncovering the community structure of customers' networks may prove to be useful for companies when planning their marketing activities.

A lot of attention has been paid to identifying community structures for a long time, and many methods have been proposed (e.g., Newman, 2006; Ng, Jordan, and Weiss, 2002; Nowicki and Snijders, 2001; Handcock, Raftery, and Tantrum, 2007). In addition to social network analysis, these methods are used in many other fields, including analysis of protein-protein interaction networks (Jeong et al., 2001), terrorists networks (Krebs, 2002), and co-author networks (Liu et al., 2005).

However, these methods focus only on network information, while more meaningful communities could be identified if other source of information was considered. For example, students belonging to the same community of “school” are thought to be connected each other to form social networks. Such networks are regarded as one community when considering only network information. At the same time, the students may be involved in various hobbies such as music, books, or sports. More meaningful segmentation can be achieved if researchers regard these networks whose members have different properties (or interests) as multiple communities rather than single community. To do so, text information on social media, or user-generated-content (UGC), can be used to uncover members’ interests.

In this study, we propose a model for identifying and characterizing online communities where not only edge structure on the network but also topics of text posted by community members are distinct from other communities, and we define such communities as *topic-based communities*. We note that text information used in this study indicate node-feature as like postings on social media and blogs not edge-feature such as message between nodes.

When we understand the community structure of a social network, we should consider the problem of multiple communities such as family, work, and online friends, in addition to topic-based communities. This problem is called community overlapping. In this case, when applying methods such as hard clustering, where each node is assumed to belong to a single community, the estimated network structure in this case can have a large deviation from that of the real network. The mixed membership stochastic block model (MMSB) proposed by Airoldi et al. (2008) is one of the most popular statistical generative models accommodating the community overlapping problem. In this study, the proposed model also share the same structure with MMSB allowing nodes to belong to different latent communities for each relationship. Therefore, the purpose of this study is to identify potentially overlapping topic-based communities by considering network and text information available on social media.

The rest of this chapter is organized as follows: related work is discussed in Section 1.2. The proposed model, and its inference algorithm are introduced in Section 1.3. Section 1.4 examines the simulation studies conducted to validate the main features of the proposed model and choose numbers of communities and topics. Section 1.5 presents an application of the proposed model to a real-world network, namely, Twitter. Finally, Section 1.6 provides some concluding remarks.

1.2 Literature Review

1.2.1 Identifying Communities Using Network Information

A number of models have been proposed in the literature to identify the community structure of a network. They can be divided into two approaches, deterministic algorithm and statistical models. One of the approaches using a deterministic algorithm is based on the modularity score introduced by Newman (2006), where modularity is a measure of the strength of connections within a network divided into modules; a network with high modularity forms dense connections between the nodes within modules but sparse connections between nodes in different modules. The algorithm proposed by Newman (2006) detects communities by maximizing modularity, and this algorithm is one of the most widely used methods due to its simplicity. Another approach using a deterministic algorithm is spectral clustering (Ng, Jordan, and Weiss, 2002), which is based on the eigenvalue decomposition of the graph Laplacian. The graph Laplacian is a matrix obtained by transforming the adjacency matrix, and the community structure can be clarified by applying some clustering methods such as k-means for the eigenvectors of the graph Laplacian.

The community detection methods using statistical models have been well developed in past decades, and the representative one is the stochastic block model (SBM) proposed by Wang and Wong (1987) and formulated by Snijders and Nowicki (1997) and Nowicki and Snijders (2001). The SBM assumes that when the cluster membership of each node is given, the relationship between nodes is generated according to some probability distribution such as the Bernoulli distribution. Recently, SBM has been extended by many researchers from the aspect of multiple memberships (multiple networks). One of the representative models is the MMSB by Airoldi et al. (2008), which allows each node to stochastically belong to multiple clusters. Also, Barbillon et al. (2017) and Latouche, Birmelé, and Ambroise (2011) extend SBM for multiple networks. Another stream of extension is on the dynamic characteristics of network evolving over time, and some dynamic SBMs have been proposed (e.g., Matias and Miele, 2017; Xu and Hero, 2014; Xing, Fu, and Song, 2010).

In the literature, it is known that a relationship between nodes is affected by node (or dyad, triad) specific features such as gender and age (Hoff, Raftery, and Handcock, 2002; Handcock, Raftery, and Tantrum, 2007; Krivitsky et al., 2009) as well as network structure. In this study, however, the proposed model does not consider such features because we focus on online communities. In offline settings (i.e. social network in the real world), when people try to have a relationship with someone, they can judge by considering the others personal information. On the other hand, in online settings (e.g. Twitter), they can register accounts with masked personal attributes. Hence, when they send a request of relationship to someone, the main information they can consider may be who they have relationships with and what contents they create, that is, network and text information considered in this study. However it is also valuable to extend the proposed model by taking node features into account toward a general social network model.

1.2.2 Simultaneous Modeling of Network and Other Information

The models introduced in Section 1.2.1 consider only network information (i.e., the connections between nodes). On the other hand, simultaneous modeling of network and text data is useful for a deep understanding of modern online networks such as Twitter and Facebook, because these two kinds of information allow researchers to recognize more valuable structures for companies by accommodating the detection of heterogeneous relationships and interests across a specific community that are hidden in network data. For instance, it is possible to detect a group of music lovers in a community of school, which the former is a topic-based community detected by text and the latter is a community identified by only network information.

In the literature, several studies on community identification considering network and other information, including text, have been developed. Firstly, one of the most prominent work for community detection considering other information on the network, not limited to text information, is the latent position cluster model (LPCM, Handcock, Raftery, and Tantrum, 2007) that extends the latent space model (LSM, Hoff, Raftery, and Handcock, 2002). Handcock, Raftery, and Tantrum (2007) introduce parameters for the position of nodes on the latent space and propose logistic regression model considering the latent positions and edge features for edge patterns. Also, Zanghi, Volant, and Ambroise (2010) propose a model assuming that the connectivity pattern and node features are independently explained when the node classes, that is, communities, are given. However in the case focusing on text information as node features, topic modeling, such as latent Dirichlet allocation (LDA, Blei, Ng, and Jordan, 2003), can be adequate for the generative model for text rather than the model of Zanghi, Volant, and Ambroise (2010) assuming node features to follow normal distribution.

Chang and Blei (2010) propose the relational topic model (RTM) applying topic model for node-specific text and assuming nonlinear functions of topic assignments for link between the nodes. However, in contrast to their purpose of RTM that grasps the topic structure using network information, our study aims to understand the community structure using text information.

Several studies propose topic models for understanding the community structure considering network and text information. Pathak et al. (2008) propose the community author recipient topic (CART) model that incorporates both network and text information to extract well-connected and topically meaningful communities. Furthermore, CART allows the nodes to belong to multiple communities. Also, the CART assumes textual edges, where text information appertains to edges, which is the case in e-mail networks and co-authorship networks of the papers and is different from the focus of this research. In addition, unlike the CART designed only for directed graphs, our model can handle both directed and undirected graphs. Also, Liu, Niculescu-Mizil, and Gryc (2009) proposed the topic-link LDA (TL-LDA) method that detects the community structure by considering information in a situation with textual nodes, which is similar to our research. However, this method assumes that each node has a single community membership. In addition, the probability of creating an edge between nodes is defined by the similarity of the community and topic proportion of the nodes. Hence, the probability is constant regardless of the direction of the edge and can be applied to undirected graphs only.

In a recent study, Bouveyron, Latouche, and Zreik (2018) proposed the stochastic block topic model (STBM) that extends the SBM by incorporating text information into the model and is suitable for both undirected and directed graphs. If a node belongs to community A and another node belongs to community B, the SBM handles any graph regardless of whether it is directed or not by estimating the probability separately for the cases of generating edges from A to B and from B to A. While our proposed method can handle the two types of graphs similar to the STBM, our method also overcomes the limitation of the STBM, where nodes can have only a single community membership.

Zhu et al. (2013) propose a model combining MMSB and LDA, both their purpose and model are similar to that of this study. The key difference is that communities and topics which are assigned to edges and words are assumed to follow the same distribution. On the other hand, in this study, each of them follows different distributions, which will be discussed in Sect. 1.3. In other words, Zhu et al. (2013) regard the dimensions of communities and topics as the same. In real social networks, however, communities and topics do not always correspond each other. For instance, when we consider a community whose members are interested in music and sports, one community corresponds to multiple topics. If the community is detected by the model of Zhu et al. (2013), words related to both topics, music and sports, are mixed in the words that characterize the community, and it is difficult for human to understand such characterization. In Sect. 1.3, we discuss how the proposed model deal with this limitation.

Finally, we clarify the characteristics of our model. Table 1.1 summarizes the discussed models compared by four characteristics. When comparing to the models that consider either network or text information only (such as Blei, Ng, and Jordan, 2003; Nowicki and Snijders, 2001), our model has an advantage of being able to extract well-connected and topically meaningful communities by taking both types of information into account. When comparing to the models that consider both types of

TABLE 1.1: Comparison between the proposed model and existing models

	Network	Other Information	Mixed Membership	Direction of graph
Blei, Ng, and Jordan (2003)	-	Node-text	○	-
Nowicki and Snijders (2001)	○	-	-	Both
Airoldi et al. (2008)	○	-	○	Both
Handcock, Raftery, and Tantrum (2007)	○	Edge-features	○	Both
Zanghi, Volant, and Ambroise (2010)	○	Node-features	-	Both
Chang and Blei (2010)	○	Node-text	-	Undirected
Pathak et al. (2008)	○	Edge-text	○	Directed
Liu, Niculescu-Mizil, and Gryc (2009)	○	Node-text	-	Undirected
Zhu et al. (2013)	○	Node text	○	Both
Bouveyron, Latouche, and Zreik (2018)	○	Edge-text	-	Both
This study	○	Node-text	○	Both

information, our model can be distinguished from the existing models according to the following three properties: nodes can have multiple community memberships; graphs can be both directed and undirected; text information appertains to nodes, which is the situation, where people post their own tweets toward all users on their Twitter timeline. Considering these features, we call our model the Mixed Membership Stochastic Topic Blockmodels (MMSTB).

1.3 Model

This section describes the proposed model, MMSTB, for identifying topic-based communities. Our observed data consist of the adjacency matrix A as a network information and bag-of-words collection W as a node-specific text information. In the following, we explain the process of generating these data and inference procedure employed in MMSTB.

1.3.1 Model Specification

First, we consider a directed network with D nodes. $D \times D$ adjacency matrix A represents the relationships between the nodes with their elements being $a_{ij} = 0$ (not connected) or 1 (connected). We assume that the network has no self-loops and therefore $a_{ii} = 0, \forall i$. For the relationship from node i to node j, we consider that sender i belongs to latent community $s_{ij} \in \{1, \ldots, K\}$ (K is the number of communities), while recipient j belongs to latent community $r_{ji} \in \{1, \ldots, K\}$.

$D \times D$ matrix representations of latent communities are denoted as $S = (s_{ij})$ and $R = (r_{ji})$, respectively. These sender and recipient communities are assumed to follow a categorical distribution, $s_{ij}|\eta_i \sim Categorical(\eta_i)$, $r_{ji}|\eta_j \sim Categorical(\eta_j)$, where $\eta_i = (\eta_{i1}, \ldots, \eta_{iK})^T$ is a community distribution which represents node i's community proportion, and $\sum_{k=1}^{K} \eta_{ik} = 1, \forall i$. The matrix representation of community proportions are denoted as $H = (\eta_1, \ldots, \eta_D)$. The prior distribution of H is assumed to follow a Dirichlet distribution, $\eta_i|\gamma \sim Dirichlet(\gamma)$ $(i = 1, \ldots, D)$, where $\gamma = (\gamma_1, \ldots, \gamma_K)$ is a hyperparameter.

We assume that the connection variable a_{ij} between node i to j, when s_{ij} and r_{ji} are given, follows the Bernoulli distribution that depends on the communities of the nodes. That is, $a_{ij}|s_{ij}, r_{ji}, \Psi \sim Bernoulli\left(\psi_{s_{ij}, r_{ji}}\right)$, where $\psi_{kk'}$ is a probability that an edge is generated when a sender node belongs to community k and a recipient node belongs to community k'. Let $K \times K$ matrix, $\Psi = (\psi_{kk'})$, be the matrix representation of edge probabilities. Each edge probability is assumed to follow a Beta distribution, $\psi_{kk'}|\delta_{kk'}, \epsilon_{kk'} \sim Beta(\delta_{kk'}, \epsilon_{kk'}), k, k' = 1, \ldots, K$, where δ, ϵ are hyperparameters of the $K \times K$ matrix.

Then, the conditional joint likelihood of the network information for parameters and latent variables, when the community distribution, H, is given, is

$$\begin{aligned}
&p(A, S, R, \Psi|H) \\
&= p(A|S, R, \Psi)p(S|H)p(R|H)p(\Psi|\delta, \epsilon) \\
&= \prod_{i=1}^{D} \left\{ \prod_{j=1, j\neq i}^{D} \{p(a_{ij}|s_{ij}, r_{ji}, \Psi)p(s_{ij}|\eta_i)p(r_{ji}|\eta_j)\} \right\} \times \\
&\quad \prod_{k=1}^{K} \prod_{k'=1}^{K} p(\psi_{kk'}|\delta_{kk'}, \epsilon_{kk'}). \qquad (1.1)
\end{aligned}$$

Next, we consider modeling text content. Node i creates some texts that are vectorized as M_i words ignoring the order, i.e., "bag-of-words". Node i's mth word w_{im} $(m = 1, \ldots, M_i)$ is assumed to have latent community $x_{im} \in \{1, \ldots, K\}$ and latent topic $z_{im} \in \{1, \ldots, L\}$ (L is the number of topics), as in the case of the conventional LDA model. The array representations of word communities and word topics are denoted as X, and Z, respectively, and each component of the arrays is a M_i-dimensional vector. We assume that word community x_{im} follows a categorical distribution, $x_{im}|\eta_i \sim Categorical(\eta_i)$. We note that η_i is a parameter for generating not only word community x_{im} but also node communities s_{ij} and r_{ij} as mentioned before, that is, η_i is a common parameter for modeling networks and texts that connects the two types of information.

A word topic z_{im} is assumed to follow a categorical distribution, $z_{im}|x_{im}, \Theta \sim Categorical(\theta_{x_{im}})$, where $\theta_k = (\theta_{k1}, \ldots, \theta_{kL})^T$ is the topic distribution representing community k's topic proportion, and $\sum_{l=1}^{L} \theta_{kl} = 1, \forall k$. The matrix representations of topic proportions are denoted as $\Theta = (\theta_1, \ldots, \theta_K)$. Each topic distribution is assumed to follow a Dirichlet distribution, $\theta_k|\alpha \sim Dirichlet(\alpha)$ $(k = 1, \ldots, K)$, where $\alpha = (\alpha_1, \ldots, \alpha_L)$ is a hyperparameter.

When a word topic z_{im} is given, the corresponding word $w_{im} \in \{1, \ldots, V\}$ is assumed to follow a categorical distribution that depends on word topic, i.e., $w_{im}|z_{im}, \Phi \sim Categorical(\phi_{z_{im}})$, where $\phi_l = (\phi_{l1}, \ldots, \phi_{lV})^T$ (V is the number of unique words in the corpus) is the word distribution representing the word generation probability, and $\sum_{v=1}^{V} \phi_{lv} = 1, \forall l$. The matrix representation of word distributions is denoted as $\Phi = (\phi_1, \ldots, \phi_L)$. Each word distribution is assumed to follow a Dirichlet distribution, $\phi_l|\beta \sim Dirichlet(\beta)$ $(l = 1, \ldots, L)$, where β is a hyperparameter.

Then, the conditional joint likelihood of text information, when H is given, is

$$\begin{aligned}
&p(W, X, Z, \Theta, \Phi|H) \\
&= p(W|Z, \Phi)p(Z|X, \Theta)p(X|H)p(\Theta|\alpha)p(\Phi|\beta) \\
&= \prod_{i=1}^{D} \left\{ \prod_{m=1}^{M_i} \{p(w_{im}|z_{im}, \Phi)p(z_{im}|x_{im}, \Theta)p(x_{im}|\eta_i)\} \right\} \times \\
&\prod_{k=1}^{K} p(\theta_k|\alpha) \prod_{l=1}^{L} p(\phi_l|\beta). \qquad (1.2)
\end{aligned}$$

Under the assumption of conditional independence of Equations (1.1) and (1.2), when nodes' community distribution, H, is given, the full joint likelihood of MM-STB is obtained by the product of Equations (1.1) and (1.2) multiplied by the density of H, $p(H|\gamma)$,

$$\begin{aligned}
&p(A, W, S, R, X, Z, H, \Psi, \Theta, \Phi) \\
&= \prod_{i=1}^{D} \left\{ \prod_{j=1, j \neq i}^{D} \{ p(a_{ij}|s_{ij}, r_{ji}, \Psi) p(s_{ij}|\eta_i) p(r_{ji}|\eta_j) \} \times \right. \\
&\qquad \left. \prod_{m=1}^{M_i} \{ p(w_{im}|z_{im}, \Phi) P(z_{im}|x_{im}, \Theta) p(x_{im}|\eta_i) \} \right\} \times \\
&\prod_{i=1}^{D} p(\eta_i|\gamma) \prod_{l=1}^{L} p(\phi_l|\beta) \prod_{k=1}^{K} \left\{ p(\theta_k|\alpha) \prod_{k'=1}^{K} p(\psi_{kk'}|\delta_{kk'}, \epsilon_{kk'}) \right\}. \qquad (1.3)
\end{aligned}$$

Here, we clarify the difference between the proposed model and the work of Zhu et al. (2013) and show how two kinds of information, network and text, helps to find topic-based communities through the comparison of two models. These two models have similar structures because both models assume MMSB for network generation and LDA for text generation and combine them. The key difference is that in their model, latent communities (s_{ij} and r_{ji} in our model) and latent topics (z_{im}) follow the same distribution (corresponding to η_i). As mentioned in previous section, using such model, it can be difficult to estimate clear and meaningful topics from networks where a single community corresponds to multiple topics. On the other hand, in our model, latent communities are generated according to community distributions and latent topics are generated according to topic distributions, and each community corresponds to topic distributions representing the proportion of topics in which community members create text. Therefore, we can consider the situation that people belong to multiple communities and the people in each community post text contents with multiple topics. In this sense, topic-based community can be defined as a community that potentially overlap with other communities and have distinct edge probabilities and topic distributions from other communities.

1.3.2 Conditional Posterior Distributions and Parameter Estimation

Many methods for estimating topic models have been proposed (e.g., the variational Bayesian method and sequential learning method). Among them, the most widely used method is the collapsed Gibbs sampler (CGS) proposed by Griffiths and Steyvers (2004), which samples only latent variables by integrating out parameters. CGS can estimate topic models more efficiently compared to the Gibbs sampler that directly samples all parameters. This study uses CGS for estimating MMSTB's parameters.

MMSTB has four types of model parameters: namely, community distributions H, edge probabilities Ψ, topic distributions Θ, and word distributions Φ. We can derive the full conditional posterior for each parameter according the conjugacy. The derivation is given Appendix.

Also, MMSTB has four types of latent variables: two latent variables for a relationship between node i and j, s_{ij} (sender community) and r_{ji} (recipient community), and two latent variables for a mth word of node i, x_{im} (word community) and z_{im} (word topic). The conditional posterior distributions of these four latent variables are derived by integrating out parameters (H, Ψ, Θ, Φ) as follows:

$$\begin{aligned}
&P(s_{ij} = k, r_{ji} = k' | a_{ij}, A_{\backslash ij}, S_{\backslash ij}, R_{\backslash ji}, X, \gamma, \delta, \epsilon) \\
&\propto \int\int P(s_{ij} = k|\eta_i) P(r_{ji} = k'|\eta_j) P(x_i|\eta_i) P(x_j|\eta_j) \\
&\quad P(\eta_i | S_{\backslash ij}, R_{\backslash ji}, X, \gamma) P(\eta_j | S_{\backslash ij}, R_{\backslash ji}, X, \gamma) d\eta_i d\eta_j \times \\
&\quad \int P(a_{ij}|\psi_{kk'}) P(\psi_{kk'} | A_{\backslash ij}, S_{\backslash ij}, R_{\backslash ji}, \delta, \epsilon) d\psi_{kk'} \\
&= \frac{N_{ik\backslash ij} + M_{ik} + \gamma_k}{\sum_t \left(N_{it\backslash ij} + M_{it} + \gamma_t\right)} \times \frac{N_{jk'\backslash ji} + M_{jk'} + \gamma_{k'}}{\sum_t \left(N_{jt\backslash ji} + M_{jt} + \gamma_t\right)} \times \\
&\quad \frac{\left(n^{(+)}_{kk'\backslash ij} + \delta_{kk'}\right)^{\mathbb{I}(a_{ij}=1)} \left(n^{(-)}_{kk'\backslash ij} + \epsilon_{kk'}\right)^{\mathbb{I}(a_{ij}=0)}}{n^{(+)}_{kk'\backslash ij} + n^{(-)}_{kk'\backslash ij} + \delta_{kk'} + \epsilon_{kk'}}
\end{aligned} \tag{1.4}$$

$$\begin{aligned}
&P(x_{im} = k, z_{im} = l | W, S, R, X_{\backslash im}, Z_{\backslash im}, \alpha, \beta, \gamma) \\
&\propto \int P(s_i, r_i|\eta_i) P(x_{im} = k|\eta_i) P(\eta_i | S, R, X_{\backslash im}, \gamma) d\eta_i \times \\
&\quad \int P(z_{im} = l|\theta_k) P(\theta_k | X_{\backslash im}, Z_{\backslash ,im}, \alpha) d\theta_k \times \\
&\quad \int P(w_{im} = v|\phi_l) P(\phi_l | W_{\backslash im}, Z_{\backslash im}, \beta) d\phi_l \\
&= \frac{M_{lv\backslash im} + \beta_v}{\sum_u \left(M_{lu\backslash im} + \beta_u\right)} \times \frac{M_{kl\backslash im} + \alpha_l}{\sum_q \left(M_{kq\backslash im} + \alpha_q\right)} \times \frac{N_{ik} + M_{ik\backslash im} + \gamma_k}{\sum_t \left(N_{it} + M_{it\backslash im} + \gamma_t\right)},
\end{aligned} \tag{1.5}$$

where the symbol $\setminus$ represents the exclusion of an edge or a word from the count number.

The algorithm of CGS for MMSTB is provided in the Appendix A.1. In CGS, according to Equations (1.4) and (1.5), the latent community and topic for each edge and word are sampled. Finally, using the samples of the latent variables excluding the burn-in samples, model parameters are point estimated.

1.4 Numerical Experiments

This section described the numerical experiments we conducted to highlight the main features of the proposed approach and provide the validity of our inference algorithm.

1.4.1 Experimental Settings

The main features of our modeling are the mixed membership of nodes and simultaneous modeling of network data and text content. The characteristic of a mixed membership captures the situation of people belonging to multiple communities on a social network and building relationships with other members of these communities. Furthermore, it is possible to extract more meaningful segments from social networks by considering both network data and text content.

To highlight these two properties of MMSTB, we have designed three different scenarios for numerical experiments. Table 1.2 provides the settings of each scenario, while Figure 1.1 depicts an example of the generated adjacency matrix, where black (white) cells mean the presence (absence) of a relationship between two nodes. We set some values for the community distribution, edge probability, and topic distribution but did not set any values for the word distribution. Instead, for all scenarios, 150 words are sampled per node according to their word topics from the BBC news document dataset (Greene and Cunningham, 2006) as virtual text contents; this dataset contains three topics: namely, business, entertainment, and sports.

Scenario A

The network and text content are composed of $K = 3$ communities and $L = 2$ topics. Each node belongs to only one community (Node 1-20, 41-60, 81-100) or two communities (Node 21-40, 61-80); that is, these communities are overlapping. But the edge probabilities across the communities are lower ($\psi_{kk'} = 0.1$) than within the communities ($\psi_{kk} = 0.5$), and each community has a unique topic proportion ($\theta_1 \neq \theta_2 \neq \theta_3$). Therefore, both MMSTB and other models using only one source of information such as LDA and MMSB can be expected to detect these communities accurately.

Scenario B

Similar to scenario A, each node belongs to one or two communities, and $K = 4$ communities are overlapping. Unlike scenario A, the community 1 and 4 have the same topic proportions ($\theta_1 = \theta_4$). Therefore, the models using only text content information cannot distinguish between the nodes that belong to only community 1 (Node 1-20) or community 4 (Node 91-100). Conversely, the edge probabilities across the communities are low; hence, both MMSTB and models using only network information should be able to distinguish all communities.

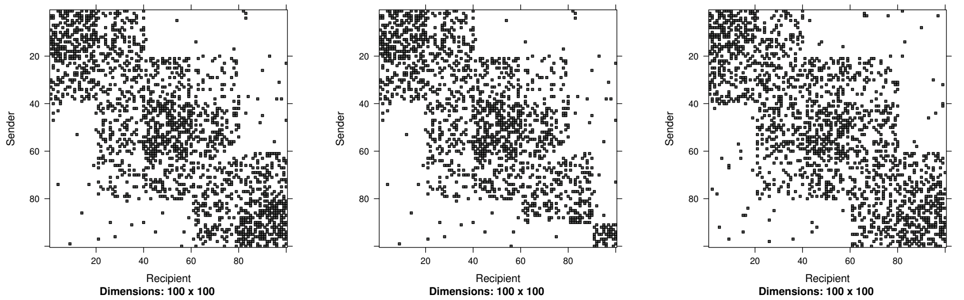
Scenario C

The community 1 and 4 have the same topic proportion, and the text content-based models cannot distinguish between these two communities. Furthermore, the edge probabilities between communities 3 and 4 (ψ_{34}, ψ_{43}) are high; that is, people in these communities are well-connected even if they have different interests (topics). Therefore, the network-based models cannot identify these two communities. Only MMSTB can detect all communities and recover the community structure properly.

We note that nodes are divided into some clusters where they belong to the same community (communities) with the same proportion and generate virtual texts of the same topic(s). Each row of H in Table 1.2 corresponds to each cluster, and, for example, in scenario A, nodes 1-20 are classified into the same cluster. Whether models can recover these cluster structures depends on the situation of each scenario as described above. In the next section, we validate whether our model and the models that are popular in the literature, namely, LDA as a text-based model and MMSB as a network-based model, are able to correctly estimate parameters and identify true cluster structures.

TABLE 1.2: The settings of three simulation scenarios

	Scenario A	Scenario B	Scenario C
D (nodes)	100	100	100
K (communities)	3	4	4
L (topics)	2	2	3
	$\{\eta_1,\dots,\eta_{20}\}:\ (1,0,0)$	$\{\eta_1,\dots,\eta_{20}\}:\ (1,0,0,0)$	$\{\eta_1,\dots,\eta_{20}\}:\ (1,0,0,0)$
	$\{\eta_{21},\dots,\eta_{40}\}:\ (.5,.5,0)$	$\{\eta_{21},\dots,\eta_{40}\}:\ (.5,.5,0,0)$	$\{\eta_{21},\dots,\eta_{40}\}:\ (.5,.5,0,0)$
Community dist.	$\{\eta_{41},\dots,\eta_{60}\}:\ (0,1,0)$	$\{\eta_{41},\dots,\eta_{60}\}:\ (0,1,0,0)$	$\{\eta_{41},\dots,\eta_{60}\}:\ (0,1,0,0)$
H	$\{\eta_{61},\dots,\eta_{80}\}:\ (0,.5,.5)$	$\{\eta_{61},\dots,\eta_{80}\}:\ (0,.5,.5,0)$	$\{\eta_{61},\dots,\eta_{80}\}:\ (0,.5,.5,0)$
	$\{\eta_{81},\dots,\eta_{100}\}:\ (0,0,1)$	$\{\eta_{81},\dots,\eta_{90}\}:\ (0,0,1,0)$	$\{\eta_{81},\dots,\eta_{90}\}:\ (0,0,1,0)$
		$\{\eta_{91},\dots,\eta_{100}\}:\ (0,0,0,1)$	$\{\eta_{91},\dots,\eta_{100}\}:\ (0,0,0,1)$
	$\theta_1 = (1,0)$	$\theta_1 = (1,0)$	$\theta_1 = (.5,0,.5)$
Topic dist.	$\theta_2 = (.5,.5)$	$\theta_2 = (.5,.5)$	$\theta_2 = (.5,.5,0)$
Θ	$\theta_3 = (0,1)$	$\theta_3 = (0,1)$	$\theta_3 = (0,1,0)$
		$\theta_4 = (1,0)$	$\theta_4 = (.5,0,.5)$
Edge prob.	$\psi_{11},\psi_{22},\psi_{33} = .5$	$\psi_{11},\psi_{22},\psi_{33},\psi_{44} = .5$	$\psi_{11},\psi_{22},\psi_{33},\psi_{34},\psi_{43},\psi_{44} = .5$
Ψ	*otherwise* .1	*otherwise* .1	*otherwise* .1



Scenario A Scenario B Scenario C

FIGURE 1.1: Adjacency matrix for each scenario

1.4.2 Reproducibility of Parameters and Recovery of Cluster Structures

This section presents the experiments we conducted to verify whether the considered models (LDA, MMSB, and MMSTB) can reproduce parameters and recover cluster structures as described in the previous section. The modeling assumptions for LDA and MMSB are taken from the original papers (Blei, Ng, and Jordan, 2003; Airoldi et al., 2008), while the generative process of these models is outlined in the supplementary material. As LDA is a model for text content and MMSB is a model for network data, we provide only text data of the simulated dataset for LDA, only network data for MMSB, and the entire dataset for MMSTB. Similar to MMSTB, we use CGS to estimate parameters of LDA and MMSB. The number of iterations is set to 5,000, and the first 2,000 samples are excluded as burn-in samples. The values of the hyperparameters for the respective prior distributions are listed in Table 1.3.

First, we carry out an experiment to verify the reproducibility of parameters. Figure 1.2 and Table 1.4 show the results of scenario C estimating MMSTB. The three panels in Figure 1.2 show the estimated parameters, community distribution (left), topic distribution (top-right), and edge probability (right-bottom). The results show that MMSTB reproduces the values provided in Table 1.2 with high accuracy. Table 1.4 lists the top 10 words for each topic in descending order of the estimated word distribution values. From left to right, words related to business, entertainment, and sports are lined up, which implies that MMSTB extracts all topics correctly. Therefore, MMSTB appropriately detects meaningful communities by allowing nodes to have mixed memberships and considering network data and text content simultaneously. The results of the other scenarios and models are provided in the supplementary material.

Next, we conducted an experiment to demonstrate the recovery of cluster structures from the simulated dataset. These cluster structures can be found using the estimated node-specific parameters. In particular, MMSB and MMSTB have a node-specific community distribution, whereas LDA has a node-specific topic distribution. For example, MMSTB's community distribution affects the generation of both the network and text data as explained in Section 1.3, while the nodes having similar values for the node-specific parameter (e.g., nodes 1-20 in scenario A have the same value for community distribution) should generate similar network and text data. Therefore, it is natural that these nodes are classified into the same cluster. In this experiment, we apply a clustering method, k-means, to the estimated node-specific parameter of each model and compare the clustering results with the true labels listed in Table 1.2.

The process of the experiment is as follows. First, we simulate datasets for each scenario according to Table 1.2. Second, we estimate the model parameters while providing text data for LDA, network data for MMSB, and both datasets for MMSTB. The number of iterations and hyperparameter values are the same as described above. Next, we classify nodes according to the estimated node-specific parameters using k-means method. Then, we calculate the adjusted Rand index (ARI, Hubert and Arabie, 1985) between the estimated cluster and true labels, with higher ARIs representing higher similarity between these labels (when the labels perfectly match, ARI is 1). Because the k-means method depends on the initial value, we independently calculate ARIs for 20 different initial values and select maximum ARI value. Finally, we repeat this process 50 times with the different seed value in generating dataset.

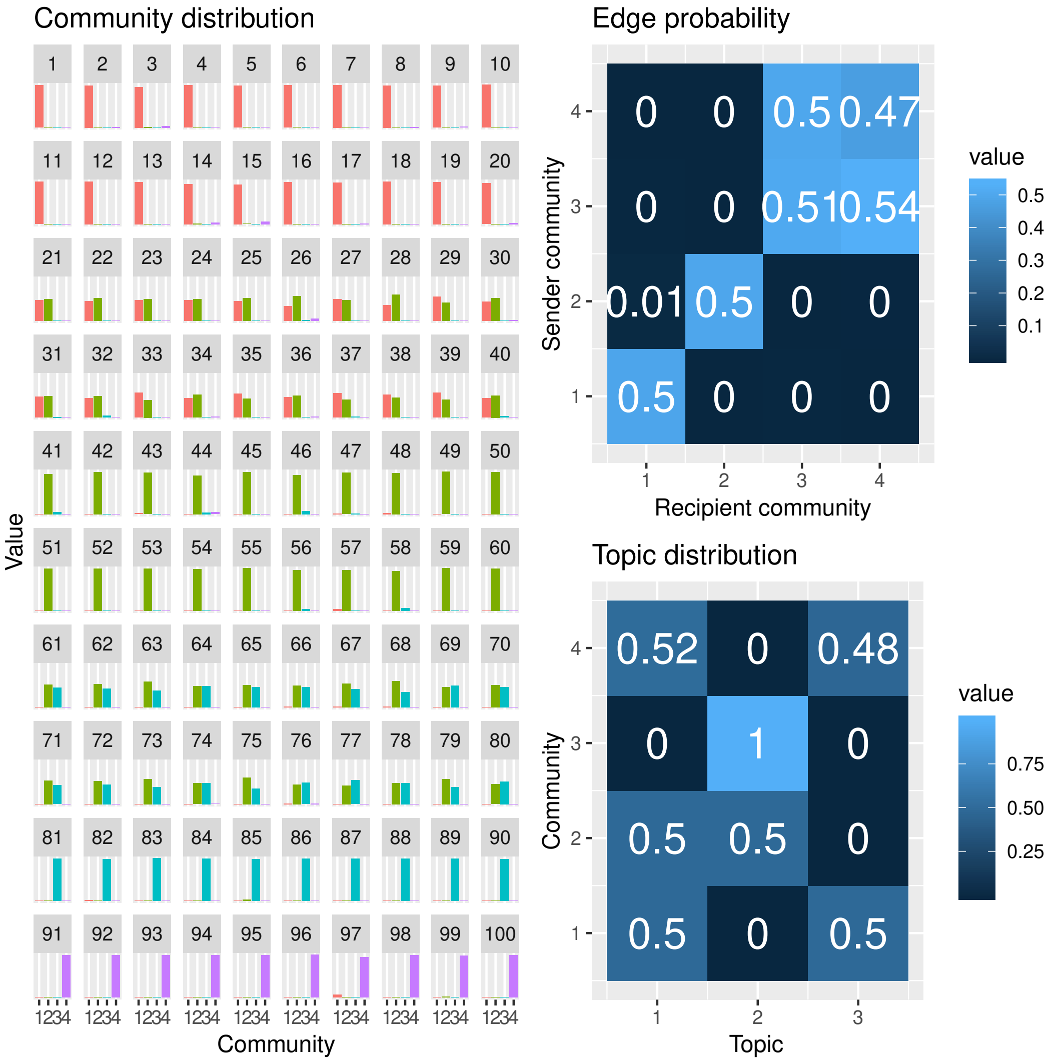
Table 1.5 lists the medians of 50 ARIs calculated for three models and three scenarios. According to the result of scenario A (first column), all medians of ARIs are 1.0; that is, all models can recover the true clusters. This result can be explained by the fact that the links within (between) communities are dense (sparse) while the topics of texts within a community are distinct from that of other communities. Even if only network data or text information are employed, differences between clusters can be identified.

According to the result of scenario B (second column), ARIs of MMSB and MMSTB are still high, whereas LDA's ARI is lower than before. In scenario B, community 1, to which nodes 1-20 (cluster 1) belong, and community 4, to which nodes 91-100 (cluster 6) belong, have the same value of the topic proportion; therefore, these clusters cannot be distinguished when looking at text data only. Conversely, non-diagonal elements of the edge probability are low; that is, the difference between these clusters is clear when considering network data. This is the reason why MMSB and MMSTB are able to recover the true clusters.

Finally, according to the result of scenario C, ARI is 1.0 only for MMSTB, whereas the ARI values of LDA and MMSB are far less than 1.0; that is, the latter models are unable to correctly cluster nodes. The reason for this result is that the text data in scenario C have the same topic structures as that of scenario B (topic distributions of communities 1 and 4 are the same). Furthermore, the edge probabilities between communities 3 and 4 are equal to the probabilities within these communities; that is, both communities completely overlap in the network. Therefore, these communities cannot be identified when considering network data only. On the other hand, MMSTB takes both network and text data into account and hence is able to recover the true cluster structures. This numerical experiment reveals that our proposed model can correctly identify structures of communities and topics even if these structures overlap, which is one of the most notable features of our model.

TABLE 1.3: The setting of hyperparameters for the simulation experiments

Hyper parameters	Prior distributions	Values
γ	$\eta_i \sim Dir(\gamma)$	$\gamma_k = 1.0, \forall k$
δ	$\psi_{kk'} \sim Beta(\delta_{kk'}, \epsilon_{kk'})$	$\delta_{kk'} = 0.1, \forall k, k'$
ϵ		$\epsilon_{kk'} = 0.1, \forall k, k'$
α	$\theta_k \sim Dir(\alpha)$	$\alpha_l = 0.1, \forall l$
β	$\phi_l \sim Dir(\beta)$	$\beta_v = 0.1, \forall v$


Community distribution

Edge probability

Topic distribution

FIGURE 1.2: The estimation results of scenario C

TABLE 1.4: Top 10 words in descending order of the word distribution for each topic

Topic 1 (Business)	Topic 2 (Entertainment)	Topic 3 (Sports)
bank	film	champion
growth	award	cup
oil	actor	coach
profit	album	rugbi
euro	nomin	ireland
stock	band	season
yuko	song	injuri
investor	oscar	olymp
award	chart	championship
deficit	actress	goal

TABLE 1.5: Medians of the adjusted Rand indices for the three models in the three scenarios

	Scenario A	Scenario B	Scenario C
LDA	1.0	0.86	0.86
MMSB	1.0	0.97	0.91
MMSTB	1.0	1.0	1.0

1.4.3 Choosing Number of Communities and Number of Topics

The numbers of communities K and topics L need to be fixed before applying SBM and its extended models (DCSB, MMSB, MMSTB, etc.). A variety of approaches has been proposed in the literature for choosing these numbers, including information criteria such as BIC (Handcock, Raftery, and Tantrum, 2007; Saldaña, Yu, and Feng, 2017), integrated completed likelihood (Daudin, Picard, and Robin, 2008; Bouveyron, Latouche, and Zreik, 2018), cross-validation (Chen and Lei, 2018), and Bayesian inference (Latouche, Birmelé, and Ambroise, 2012; McDaid et al., 2013).

In this study, the numbers of communities and topics are determined using an information criteria based on its solid theoretical ground and convenience of calculating from the outputs of CGS. However, the topic models (including MMSTB), which have latent variables, are known as singular models, and information criteria for regular models such as AIC and BIC are not appropriate. Therefore, we employ the widely applicable information criterion (WAIC Watanabe, 2010) because it can be applied to both regular and singular models. WAIC estimates the expected pointwise predictive density for a new dataset. It is defined as $-2(lppd - p_{waic})$, where

TABLE 1.6: The number of times WAIC selects each MMSTB model (K, L) in 50 simulations of each of the three scenarios

Scenario A ($K = 3$, $L = 2$)							Scenario B ($K = 4$, $L = 2$)							Scenario C ($K = 4$, $L = 3$)						
		Topics (L)							Topics (L)							Topics (L)				
		2	3	4	5	6			2	3	4	5	6			2	3	4	5	6
Communities (K)	2	0	0	0	0	0	Communities (K)	2	0	0	0	0	0	Communities (K)	2	0	0	0	0	0
	3	46	2	0	0	0		3	0	0	0	0	0		3	0	0	0	0	0
	4	2	0	0	0	0		4	38	1	0	0	0		4	0	46	1	1	0
	5	0	0	0	0	0		5	11	0	0	0	0		5	0	2	0	0	0
	6	0	0	0	0	0		6	0	0	0	0	0		6	0	0	0	0	0

$lppd$ denotes the log pointwise predictive density representing the predictive accuracy of the fitted model to data, and p_{waic} denotes a term to correct for bias due to overfitting[1]. The definition of WAIC for MMSTB is provided in the Appendix B.1.

In addition to the reproducibility of MMSTB parameters of described above, we confirm that the numbers of communities and topics can be correctly estimated by the model selection using WAIC. The procedure of the model selection simulation is as follows. For each scenario, we generate simulation data according to the values listed in Table 1.2. We estimate the models within the range of numbers of communities and topics from 2 to 6, and the model with the smallest WAIC is selected. The results of repeating these procedures 50 times are shown in Table 1.6. In all three scenarios, the model selection using WAIC succeeds in identifying the correct combination of the numbers of communities and topics. These experiments allow us to validate WAIC as a model selection criterion for MMSTB.

1.5 Empirical Analysis

1.5.1 Dataset

In this section, we apply our model to empirical data to demonstrate the usefulness of MMSTB for actual online networks. In particular, we employ the Twitter platform and user-generated text data collected by the authors. We focus on a Twitter ego network centered on the official account (@NintendoAmerica) operated by a subsidiary company of Nintendo Co., Ltd. in U.S., Nintendo of America Inc. We created a dataset for analysis according to the following procedure.

[1] In this study, we use the Gelman et al. (2013)'s scale with $-2n$ times Watanabe (2010)'s original definition (n is the number of data). This scale enables us to compare with other information criterion such as AIC and DIC

TABLE 1.7: WAIC of each model of MMSTB estimated for the Twitter dataset

		Topics (L)					
		5	6	7	8	9	10
Communities (K)	5	4,601,682	4,591,215	4,547,102	4,651,380	4,651,888	4,521,875
	6	4,633,828	4,580,391	4,564,193	4,568,752	4,629,114	5,563,824
	7	4,607,615	4,588,504	4,627,986	4,564,135	4,596,299	4,553,339
	8	4,613,074	4,637,185	4,623,877	4,517,891	4,564,046	4,537,160
	9	4,612,382	4,626,961	4,557,745	4,540,766	**4,500,094**	4,571,307
	10	4,598,036	4,580,622	4,580,856	4,544,071	4,534,666	6,629,801

First, users were randomly sampled from the users who follow the official account of Nintendo of America based on the following-followed relationship on May 1, 2018. Next, additional users were randomly sampled from the users who follow the users following the Nintendo account. The users whose average of the numbers of followers and followees is less than 3 in this network were excluded as outliers (note that the numbers of followers and followees are the numbers in the dataset and not the actual numbers). As a result, the number of selected users is 3,500, the number of total link edges are 68,949 (i.e., each user has 19.7 edges on average), and their directed relationships are used as network information.

Next, we collected the tweets posted by the selected users on their timelines from September 1, 2017 to February 28, 2018 [2]. These tweet data were preprocessed as follows: decomposing into word sets for each user, changing to lowercase letters, excluding numbers, symbols, and some popular stop-words (a, the, I, etc.) and reducing inflected words to their word stem. Among the preprocessed words, we excluded those with low frequencies (words having the number of occurrences in the corpus less than 20 or used by less than 20 users) or high frequencies (words used by more than 50 users) because these words may adversely affect the topic extraction. Then, the users whose number of words is less than five are also excluded. As a result, the number of unique words in the corpus is 9,001, and the average number of words per node is 98.2 (the average unique word number is 59.3). Next, we applied MMSTB to this Twitter dataset. The model selected by WAIC was $(K, L) = (9, 9)$ as shown in Table 1.7.

[2] We confirmed that the majority of users posted about the presentation of a new game software, called Nintendo Direct, in March 2018. Hence, in this study, to avoid the effect of such text information commonly posted by many users, we decided to limit the period of data to be until February 28, 2018.

1.5.2 Empirical Results

In this section, we discuss about the estimated results. First, interpreting the meaning of each topic is necessary to understand what kind of interest people in the community display. Figure 1.3 shows the top 10 words for each topic. The meaning of topics and their related words are as follows: topic 1 is animation topic (e.g., blackclover, hunter×hunter, and jojos_bizarre_adventure are the titles of animations); topic 2 is game topic (e.g., steinsgate, xenovers, and acnl, Animal Crossing: New Leaf, are the titles of game software); topic 3 is e-sports topic (e.g., hori and mkleosaga are words related to fighting-games, while wnf and mdva are e-sports specific words); topic 4 is music topic (e.g., vevo, spinrilla, and wshh are websites for music); topic 5 is everyday life topic (e.g., people post texts and images of their everyday life with the hashtags of dogsoftwitter and momlife); topic 6 and 7 are business topics (e.g., digitalmarket, socialmediamarket, and contentmarket are the hashtags which are sometimes used in a business-related tweet); topic 8 is streaming and broadcasting topic (e.g., teamemmmmsi, twitchkitten, roku, and wizebot are words related to streaming or broadcasting); topic 9 is sports topic (e.g., orton and sdlive, oiler, horford, and herewego are wrestling, ice hockey, basketball, and american football specific words, respectively).

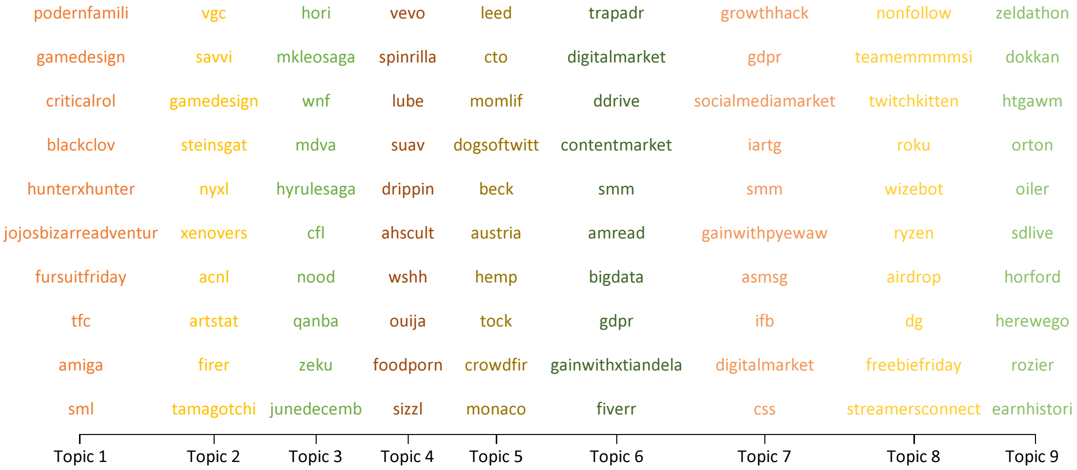
Next, Figure 1.4 shows the estimated parameters, edge probability and cumulative sum of community distributions. Looking at this figures, most of the estimated values are very low. This is because not only the network used in this study but also general social networks are very sparse, that is, few people connect to many others, while many people do not have so much connections. The estimated edge probability reflects such characteristics. But some parameters with respect to small communities, such as community 2 and 5, are estimated high, hence our model extracts small but dense network structure.

However, we can not obtain much information from the estimated result for entire network because of its large scale and sparse structure. Therefore, we look into local sub-graph structure. The interpretation of a huge network, such as these Twitter data, is hardly achievable even if we looked at the entire network image. However, the local sub-network and the estimated parameters corresponding to them provide useful some insights, in this study, on the relationship between nodes, overlapping communities, their proportions of belonging communities, and characteristic topics within each community. In Figure 1.5, the bar-graphs in circles show the values of the node's community distribution, η_i; the bar-graphs surrounding network are the values of the community's topic distribution, θ_k; and arrows represent that there is a following relationship between the nodes, where the start node of the arrow is a sender of the following, the end node of the arrow is a recipient of the following, while the bi-directed arrow means the mutual-following relationship. As an example, nodes 95 and 336 belong to community 5, in which people often post sports-related tweets (Topic 9). Node 95 belongs to not only community 5 but also community 1 related to music (Topic 4) together with other nodes (804, 2241, 3476). Thus, the communities detected by our model represent a subset of nodes with not only dense links on the network but also similar topics in their texts and overlap each other.

In the field of consumer behavior analysis, researchers know that product or information diffusion tends to become faster among people located in a well-connected area of their social network (i.e., people in the same community) as discussed in Muller and Peres (2019). In addition to the network effect, people in the community identified by our model share their interests owing to the same topic of texts posted on Twitter. Therefore, our model can help companies to detect some useful community structures that positively affect the consumption behaviors. By analyzing the relationship between a company's followers and their text content using our model, companies and managers can understand the community structures and the interests of the customers connected through these communities. Then they can use the obtained knowledge to update their marketing strategies accordingly.

1.5.3 Predicting on Holdout Samples

In this section, we compare the predictive performance of the proposed model with that of relevant models to demonstrate the predictive performance of these models on some test data generated by holding out a part of the dataset described in Section 1.5.1. Unlike the analysis outlined in the previous section, where the entire dataset was used for the model estimation, in this experiment, 90% of edges of each node



Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
podernfamili	vgc	hori	vevo	leed	trapadr	growthhack	nonfollow	zeldathon
gamedesign	savvi	mkleosaga	spinrilla	cto	digitalmarket	gdpr	teamemmmmsi	dokkan
criticalrol	gamedesign	wnf	lube	momlif	ddrive	socialmediamarket	twitchkitten	htgawm
blackclov	steinsgat	mdva	suav	dogsoftwitt	contentmarket	iartg	roku	orton
hunterxhunter	nyxl	hyrulesaga	drippin	beck	smm	smm	wizebot	oiler
jojosbizarreadventur	xenovers	cfl	ahscult	austria	amread	gainwithpyewaw	ryzen	sdlive
fursuitfriday	acnl	nood	wshh	hemp	bigdata	asmsg	airdrop	horford
tfc	artstat	qanba	ouija	tock	gdpr	ifb	dg	herewego
amiga	firer	zeku	foodporn	crowdfir	gainwithxtiandela	digitalmarket	freebiefriday	rozier
sml	tamagotchi	junedecemb	sizzl	monaco	fiverr	css	streamersconnect	earnhistori

FIGURE 1.3: Top 10 words in descending order of the word distribution for each topic of Twitter data

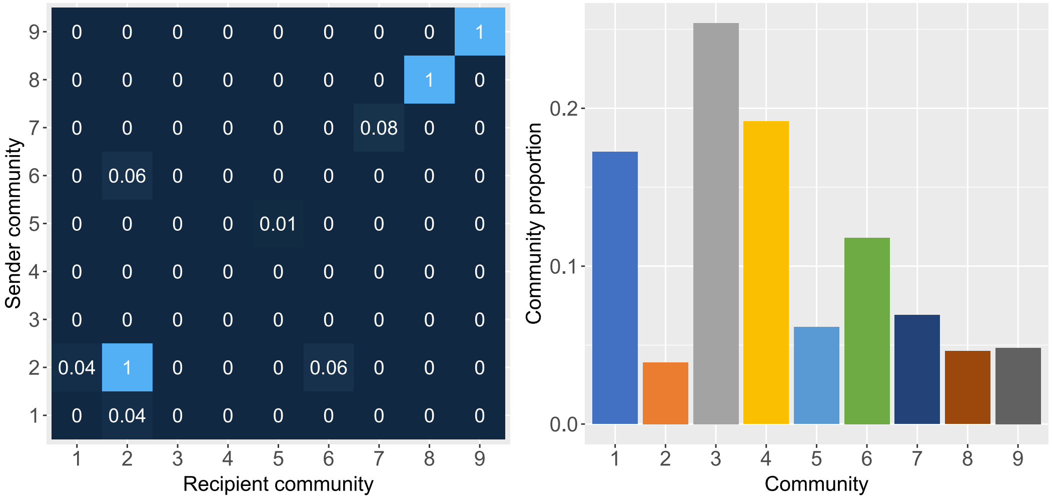


FIGURE 1.4: The estimated edge probability (left) and cumulated community distribution (right)

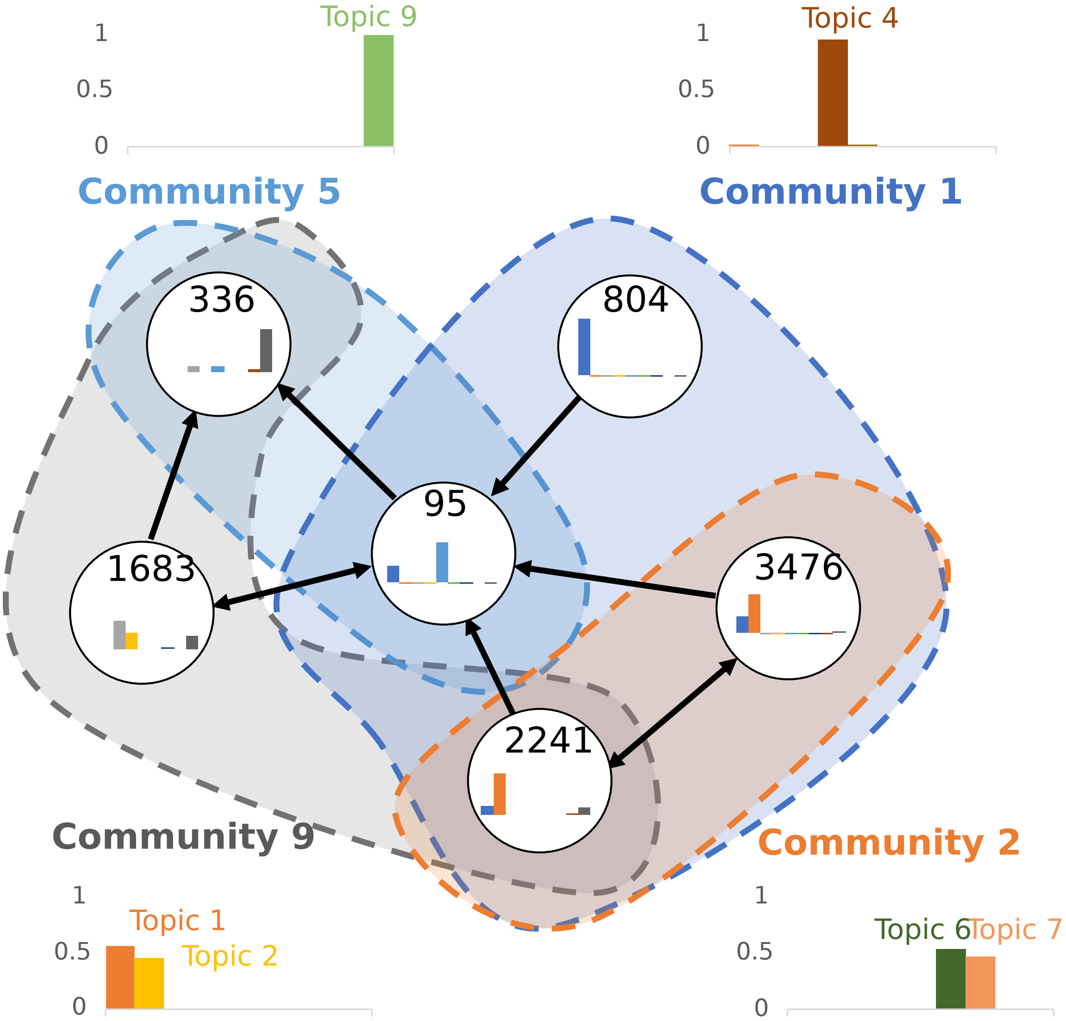


FIGURE 1.5: Sub-network consisting of a specific node (node 95) and its neighbors, and the results estimated by MMSTB

with $D-1$ edges are selected randomly as training data, while the remaining 10% of edges are used as test data. For the text data, all words of each node are used as training data. The settings of the hyperparameters are the same as listed in Table 1.3.

We consider two comparable models: the extant network model for baseline of Airoldi et al. (2008) which ignores text information on the network and the most similar model of Zhu et al. (2013) which considers network and text information on the network. Difference between Zhu et al. (2013) and the proposed model is whether latent communities for each relationship and latent topics for each word follow the same distribution or the distinct distributions. As described in Section 1.3, in the proposed model, latent communities and topics follow the community distribution and the topic distribution, respectively, while in the Zhu et al. (2013)'s model, these latent variables follow the same distribution. Therefore, a latent communities s_{ij} (sender) and r_{ji} (recipient) and a latent topic z_{im} follow the same categorical distribution, $s_{ij} \sim categorical(\eta_i)$, $r_{ji} \sim categorical(\eta_j)$, and $z_{im} \sim categorical(\eta_i)$. Since this model looks at the community structure in the network and the topic structure in the text on the same dimension, it makes a strong assumption that one community corresponds to one topic. On the other hand, the proposed model has a structure in which latent topics follow a topic distribution for each community, which is different from the community distribution followed by latent communities, so it can more flexibly capture the topic structure for each community. In Zhu et al. (2013)'s model, given latent communities and topics, the observed relationship and words are assumed to be generated from the Bernoulli distribution, $a_{ij} \mid s_{ij} = k, r_{ji} = k' \sim Bernoulli(\psi_{kk'})$, and the categorical distribution, $w_{im} \mid z_{im} = k \sim categorical(\phi_k)$, which are the same formulation with the proposed model[3].

Let be $\hat{H}$ and $\hat{\Psi}$ be the estimated community distribution and edge probability, the predictive probability of the test network data $a_{ij} \in A^{(test)}$ for the proposed model can be calculated as follows.

$$p(a_{ij} = 1) = \sum_{k=1}^{K} \sum_{k'=1}^{K} \hat{\eta}_{ik} \hat{\eta}_{jk'} \hat{\psi}_{jkk'}. \tag{1.6}$$

[3] In the original model of Zhu et al. (2013), relationship variable follows a Poisson distribution, not Bernoulli, because they assume multiple graph in which relationships can be natural integers rather than binaries. But this study assume only single graph, and we have changed the formulation as shown in the text to make a clear comparison with the proposed model.

The predictive probabilities for the another two models also can be calculated by the product of the community distribution and the edge probability in the similar way. Here, the number of communities and topics were set to from 5 to 10 (the number of topics is valid only for the proposed model) and the area under the curve (AUC) of each model was calculated in a grid. For each model and each number of communities and topics, we repeated the model estimation and the calculation of the AUC 30 times while shuffling the test data. Figure 1.6 includes box-plots summarizing the AUCs for each mode and the number of communities and topics.

Compared with (Airoldi et al., 2008) considering only network information, other two models considering both network and text information (Zhu et al., 2013, and this study) show better predictive performance on holdout samples. Furthermore, this increase is bigger as the number of communities increases. In other words, when dividing a community into smaller multiple communities, text information can help with robust segmentation to predict holdout samples even in situations where network information alone cannot be expected to provide such precise community segmentation. However, compared with Zhu et al. (2013)'s model, the proposed model slightly outperforms the predictive performance (especially the number of topics is 6 and 7), but there is no significant difference.

As a result, it can be said that the proposed model shows a significant improvement from the baseline model, but not enough improvement from the existing model, which has a similar model structure. However, the author argues that the proposed model can provide richer marketing implications from the estimation results. This difference is derived from whether the community structure and topic structure are viewed in the different dimensions or not as explained above, and in the next section, their marketing implications will be discussed in detail.

1.5.4 Marketing Implications

In the many fields of marketing, much research has studied on how various consumer behaviors, such as purchases, awareness, and learning, are influenced from the consumers' social network structure. For example, information with strong trust

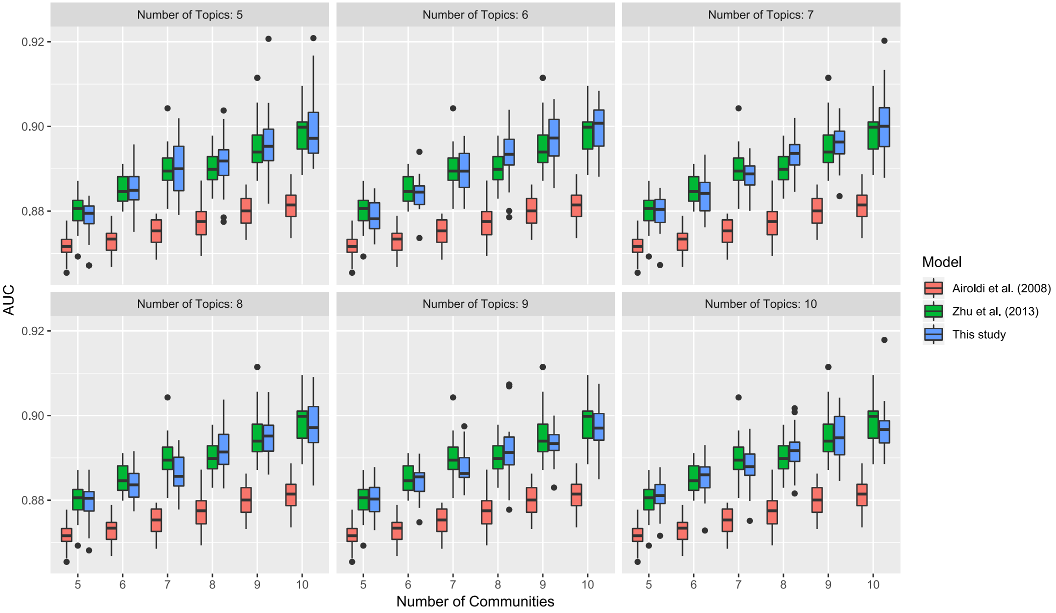


FIGURE 1.6: AUC values for comparable models and the proposed model

from social circles makes consumer's learning easier (Burt, 2001), and social interactions effectively enhance awareness in niche markets (Leskovec, Adamic, and Huberman, 2007). Especially, a number of studies have investigated the network effects on purchase behaviors or product diffusion, and the relevance literature have revealed that high average degree, that is, more ties per node, leads to faster takeoff (Delre, Jager, and Janssen, 2007; Mukherjee, 2014) and father penetration (Keeling, 2005) and is associated with faster growth (Rand and Rust, 2011) and higher net present value (Peres and Van den Bulte, 2014).

Together, these studies indicate that consumers on dense social network, where it has high average degree and social interactions are likely to occur, play a valuable role for company through awareness, learning, and purchase. Therefore, it is effective for company and managers to find the substructure such that nodes with high degree are densely located, that is, the community structure in the network. In addition, with the development of social media today, consumers are communicating with each other more actively, and companies also participate in social media, which make it easier to analyze the consumer networks around them. Companies can identify the beneficial community structure in their own customer networks to optimize various marketing activities such as management of social media accounts and planning advertising strategies.

However, when trying to propagate information or diffuse products to consumers on the social network, it is beneficial to take into account their areas of expertise and topics of interests. Muller and Peres (2019, pp. 11) state that "an opinion leader in one domain (such as whether to adopt an innovative medical treatment in a physicians' social network), might not be an opinion leader in other domains (such as adopting a new technological app in the neighborhood parents' social network)." Therefore, networks model including the proposed model and Zhu et al. (2013)'s model, which performs community extraction while capturing topic structures from the user generated text information, can be useful models for companies analyzing their own social networks.

In capturing topic structure, the difference between finding the community and topic structure in the same dimension or different dimensions is key for the marketing implications from the estimated results. Zhu et al. (2013) capture two different latent structures, communities on the network and topics in the text, with a single parameter. In other word, their assumed situation is that people within a community are interested in only one topic, but this may be far from a real. Instead, it can be a more realistic assumption that people are interested in multiple topics within a community, and the topic proportions are distinct for each community. The proposed model reflects such a situation by distinguishing between the parameters of community proportions and topic proportions.

These different assumptions are associated with the topic interpretability because when we use the model of Zhu et al. (2013) for a network where one community corresponds to multiple topics, the estimated topic corresponding that community consists of some miscellaneous topics, and it leads to less interpretability of topics. On the other hand, as it was shown that topic distributions of some communities were estimated with multiple topics in Section 1.5.2, the proposed model can identify multi-modality of people interests for each community. This allows companies and managers to understand the pair of structure of the consumer network around them and the topics consumers are interested in, so they can appropriately choose the marketing strategies, such as social media management and advertising planning, by propagating information and diffusing products according to the estimated community and topic structure.

However, we still could not significantly improve the predictive performance from the existing state of the art model of Zhu et al. (2013), and in Chapter 2, we will tackle to improve the performance by extending the edge probability formulation.

1.6 Conclusion

This study proposed a model for identifying realistic and beneficial communities based on not only the relationships within a network but also interests of its members reflected by user-generated-content, that is, topic-based communities. The main features of our model are (1) extracting communities and topics considering network information, which represents the relationships between network nodes, and text information posted on social media, which uncovers people interests, (2) allowing each node to belong to multiple communities, which is called mixed membership, and (3) being applicable to both directed and undirected graphs.

The collapsed Gibbs sampler was used for the model inference, and the numbers of communities and topics were chosen according to an information criterion. Numerical experiments using simulated data confirmed the model features and showed that the procedures of the model inference and model selection work properly. As an empirical application, we analyzed Twitter data and found realistic and beneficial community structures that could not be obtained unless both network and text information were considered. Furthermore, the proposed model demonstrated a good predictive performance on holdout samples.

In this study, we focus on the situation of online network and assume that people create relationships between others taking account with their network and text information. But in the literature of social network analysis, as another source of information influencing on the edge generation process, researchers consider node (or dyad, triad) specific features (e.g., demographic information such as gender and age). It is worthwhile to extend the proposed model for accommodating demographic features. It is possible to assume the influence of node specific features directly in the edge generation model or to treat them as a prior distribution of model parameters.

Further work may include the problem of node heterogeneity. Stochastic block model and other extended models assume that all nodes belonging to the same community are homogeneous and edges within or between communities are generated according to the corresponding community's edge probability. However, in general social networks, a few hub nodes tend to have many edges, while many other nodes tend to have a few edges, even if they belong to the same community. This property is called scale-free. Ignoring this node heterogeneity may lead to a deviation of model results from the real network structure. Krivitsky et al. (2009) introduced a parameter representing node heterogeneity for the edge generating part of the latent space model. Karrer and Newman (2011) proposed an extended model of stochastic block model that corrects the probability of generating an edge between a pair of nodes considering the node degrees to address the problem of node heterogeneity. This problem cannot be avoided in social network analysis; hence, our model also needs to be extended to solve the problem.

Chapter 2

Social Network Model Extended by Considering Node Degree Heterogeneity

2.1 Introduction

With the popularity of social networking sites (SNS) and e-commerce sites, analyzing and understanding the structure of the social network surrounding consumers have become an important part of marketing activities. The methods of social network analysis have been studied for many years, mainly in the fields of statistics and sociology, and many statistical models have been proposed to summarize and understand focal networks, such as models for extracting their community structure (e.g., Snijders and Nowicki, 1997; Airoldi et al., 2008). These models use nodes and edges in the network as observational data and extract a community structure defined as a set of nodes with a higher edge density than the surroundings. Nodes in a social network represent people, and thus, researchers have studied with the aim of refining network models by taking into account their attribute and behavior data (e.g., Handcock, Raftery, and Tantrum, 2007). In recent years, due to the popularity of SNS and e-commerce sites with functions of posting word-of-mouth, some models for social network analysis that combine user-generated-content (UGC), especially text information, with network information have been proposed (e.g., Liu, Niculescu-Mizil, and Gryc, 2009; Bouveyron, Latouche, and Zreik, 2018).

An advantage of building a model considering not only network but also text information is that it allows us to identify the community structure that is difficult to capture by either information alone. Conventional research that considers only network information defines a community as a set of nodes that have a higher edge density than others, whereas recent studies on building network model define a community by not only the density of edges but also the similarity of text information such as the topic ratio. For example, Igarashi and Terui (2020) named such communities topic-based communities and showed that a model that takes both into account allows for a more accurate division of communities than comparative models using only network or text. Therefore, estimating people's interests from generated text contents and combining them with network information not only refines social network analysis but also provides a meaningful way to understand the complex structure of contemporary online social networks.

However, another important property of social networks, besides the posted text information, is that node degree is heterogeneous from node to node, known as *node degree heterogeneity*. This property means that many people have relationships in the network with only a few people, while a limited number of people tend to have relationships with many people. Therefore, the probability of connecting edges in a social network is not constant. Moreover, it is more realistic to assume heterogeneity for each node; however, many representative network models, such as stochastic block models (e.g., Snijders and Nowicki, 1997; Igarashi and Terui, 2020) do not consider node degree heterogeneity. Thus, this study extends the model of Igarashi and Terui (2020) by making the edge generation probability different for each node, and proposes a model that takes node degree heterogeneity into account. In the empirical analysis, we use the dataset obtained from Twitter as a real social media and estimate our model for illustrating network summarization by combining network and text information with node degree heterogeneity in mind. Moreover, to verify the effects of node degree heterogeneity and the use of text information on the prediction of the out-of-sample network, we compare the proposed model with a comparative model that excludes the respective features.

The rest of this chapter is organized as follows. Section 2.2 presents the related works on social network analysis to clarify the purpose and position of this study. Section 2.3 introduces the proposed model and estimation procedure. Section 2.4

reports the empirical analysis using the Twitter dataset and predictive performance comparison. Finally, Section 2.5 provides some concluding remarks.

2.2 Literature Review

2.2.1 Progress in the Social Network Model

Researchers have studied social network modeling for a long time to understand their structure mainly in statistics and sociology. The stochastic block model (SBM, e.g., Wang and Wong, 1987; Snijders and Nowicki, 1997) is one of the most representative one. SBM assumes that nodes belong to only one of the K communities, thus, let $z_i \in \{1, \dots, K\}$ be node i's community, and then edge generation probability between nodes i and j is represented as ψ_{z_i,z_j}. This is a (z_i, z_j) element of $K \times K$ matrix Ψ, which indicates the edge probability.

SBM has been extended in various contexts. While SBM assumes a single membership for nodes, Airoldi et al. (2008) proposed a mixed membership stochastic block model (MMSB) that allows each node to belong to a different community for each relationship with other nodes. On a relationship from node i to node j, let a community node i and j belonging to be s_{ij} (sender) and r_{ji} (receiver), respectively, edge generation probability between them is represented as $\psi_{s_{ij}r_{ji}}$. This extension allows MMSB to take into account community overlapping for more realistic modeling, while overlapping community cannot be found by SBM.

In the context of sociology, the relationships between people are determined by the influence of personal characteristics such as gender and age (Hoff, Raftery, and Handcock, 2002; Handcock, Raftery, and Tantrum, 2007; Krivitsky et al., 2009). However, this study focuses on online social networks, like Twitter; hence, we do not consider such characteristics because on such anonymous social media allowing users to register an account while hiding personal information such as gender and age, the only considerable information when relating to others is the network they form and the (text) content they post on the media. Although this study does not extend the model to take into account such personal information, if data indicating node attributes are available, we can easily incorporate them into the proposed model and analyze them from a sociological perspective.

2.2.2 Studies on Simultaneous Modeling Network and Text Information

In the studies on social network models mentioned in the previous section, they focused only on the network information, but in recent years, models that consider both network and text information have been actively studied to better understand the structure of online social networks such as Twitter and Facebook. For example, Chang and Blei (2010) applied a topic modeling to node specific text information and proposed the relational topic model (RTM) in which the edge generation probability between nodes is defined according to the similarity of the topic proportions of the text posted by the nodes. However, the purpose of the analysis is contrasting with Igarashi and Terui (2020) and this study in that RTM takes into account network information to estimate topics in the text, whereas we take into account text information to understand the community structure.

Similar to Chang and Blei (2010), several studies incorporate text information into a network model using the latent Dirichlet allocation (LDA, Blei, Ng, and Jordan, 2003) or its extended models. For example, Liu, Niculescu-Mizil, and Gryc (2009) proposed the topic-link LDA, which has the same objectives as this study in terms of detecting community structure by taking into account node specific text information. However, similar to SBM, topic-link LDA has a limited assumption that nodes have to belong to a single community. Also, the edge generation probability is defined by the community memberships and the similarity of node-specific topic proportions. Therefore, the model can only be applied to undirected graphs where the generation probability does not change even if the direction of the edge is reversed, whereas block models including the proposed model defining edge probability parameters as asymmetric matrix allow us to apply them regardless of the direction of the graph. Elsewhere, Bouveyron, Latouche, and Zreik (2018) proposed the stochastic topic block model, which is an extension to the SBM in the form of adding a model for text information.

While above models extend SBM assuming a single membership, Zhu et al. (2013) proposed a network model that assumes mixed membership for nodes and considers both network and text information. The difference with the proposed model is that in the model of Zhu et al. (2013), communities assigned to a pair of nodes and topics assigned to words are assumed to be generated by following the same distribution; in other words, they treat the dimensions of community and topic as the same. However, in real world social networks, communities and topics do not always correspond to each other. Consider a network where people with interests in music and sports exist within the same community. If such community is detected by Zhu et al. (2013)'s model, a single topic with multiple semantic coherent, music and sports, correspond to one community, then the topics lack interpretability. Meanwhile, Igarashi and Terui (2020) and this study assume that communities and topics follow a different distribution. Thus, for the above network, we can map those topics to the community while recognizing music and sports topics as separate topics.

Based on these existing models, we extend the simultaneous modeling with network and text by Igarashi and Terui (2020), which assume node degree homogeneity, to consider a model in which the edge generation probability is heterogeneous for each node. This allows modeling that takes into account node degree heterogeneity that real social network typically have. In the literature, Karrer and Newman (2011) considered node degree heterogeneity by introducing the expected degree of each node into a network generative model and correcting the edge generation probability according to the extent of the corresponding nodes' heterogeneity. Meanwhile, the proposed model directly estimates the edge generation probability itself as a heterogeneous parameter for each node.

2.3 Model

2.3.1 Model Specification

In this section, we explain how this study constructs a social network model while taking into account node degree heterogeneity, while identifying the differences with Igarashi and Terui (2020), which is the basis of the proposed model. First, we describe the model of Igarashi and Terui (2020), but we will not go into details because this has already been introduced in Chapter 1.

Recalling that in Chapter 1, it was assumed that each node has a unique community distribution $H = (\eta_1, \ldots, \eta_D)$ and the network is generated by the edge generation probability Ψ corresponding to the latent communities, the conditional likelihood for the network data, given the community distribution H, is defined as follows.

$$\begin{aligned}
&p(A, S, R, \Psi \mid H) \\
&= p(A \mid S, R, \Psi) p(S \mid H) p(R \mid H) p(\Psi \mid \delta, \epsilon) \\
&= \prod_{i=1}^{D} \left\{ \prod_{j=1, j \neq i}^{D} \{ p(a_{ij} \mid s_{ij}, r_{ji}, \Psi) p(s_{ij} \mid \eta_i) p(r_{ji} \mid \eta_j) \} \right\} \times \\
&\quad \prod_{k=1}^{K} \prod_{k'=1}^{K} p(\psi_{kk'} \mid \delta_{kk'}, \epsilon_{kk'}),
\end{aligned} \tag{2.1}$$

where please refer to Section 1.3 for what each parameter means.

The community distribution was a common parameter in the process of generating both network and text data. Recalling that the proposed model assumed that latent topics of words were generated by the topic distribution, Θ corresponding to the latent word community according to the community distribution and words were generated by the word distribution Φ corresponding to the latent word topic, the conditional likelihood for the text data, given the community distribution H, is defined as follows.

$$\begin{aligned}
&p(W, X, Z, \Theta, \Phi \mid H) \\
&= p(W \mid Z, \Phi) p(Z \mid X, \Theta) p(X \mid H) p(\Theta \mid \alpha) p(\Phi \mid \beta) \\
&= \prod_{i=1}^{D} \left\{ \prod_{m=1}^{M_i} \{ p(w_{im} \mid z_{im}, \Phi) p(z_{im} \mid x_{im}, \Theta) p(x_{im} \mid \eta_i) \} \right\} \times \\
&\quad \prod_{k=1}^{K} p(\theta_k \mid \alpha) \prod_{l=1}^{L} p(\phi_l \mid \beta).
\end{aligned} \tag{2.2}$$

The conditional likelihood of Equations (2.1) and (2.2) is assumed to be independent given the community distribution, so the joint distribution of the model by Igarashi and Terui (2020) is obtained by multiplying the density of Equations (2.1), (2.2), and H as Equation (1.3).

Igarashi and Terui (2020) aimed to understand the community structure on network while considering user-generated textual content, i.e., to find topic-based communities. At this point, they assume that the edge generation probability between nodes is homogeneous for all nodes with $p(a_{ij} = 1 \mid s_{ij} = k, r_{ji} = k') \sim Bernoulli(\psi_{kk'})$. However, as explained in the previous section, in real social networks, the node degree varies greatly according to node characteristics such as gender and age. Thus, the method of Igarashi and Terui (2020), who did not consider the node degree heterogeneity, may not be able to adequately fit to the real-world network data.

In this study, to tackle the problem, we extend the model by setting the part of the edge generation probability as $p(a_{ij} \mid s_{ij} = k, r_{ji} = k') \sim Bernoulli(\psi_{jkk'})$, where $\psi_{jkk'}$ indicates the probability that edge is generated between sender node i belonging to community k and receiver node j belonging to community k'. Thus, it is a heterogeneous parameter depending on receivers of relationships. With this formulation, the proposed model reflects the heterogeneity of the degree distribution in social networks and allows us to construct a more realistic model for social network analysis by heterogeneously estimating the edge generation probability depending on the number of node degrees. In other words, the node degree heterogeneity can be interpreted as that the edge connectivity differs for each node even if both nodes have similar community distributions, which is exactly that the edge probability parameter represents node degree heterogeneity. Also, matrix representation of edge probability is $\Psi_i = (\psi_{ikk'})$, whose each element is assumed to follow Beta distribution as prior structure, $\psi_{ikk'} \mid \delta_{kk'}, \epsilon_{kk'} \sim Beta(\delta_{kk'}, \epsilon_{kk'})$.

The proposed model adopts the same formulation of Igarashi and Terui (2020), except for the above points. Here, the conditional likelihood for the network data given the community distribution H is changed from Equation (2.1) to the following.

$$
\begin{aligned}
& p(A, S, R, \Psi \mid H) \\
& = p(A \mid S, R, \Psi) p(S \mid H) p(R \mid H) p(\Psi \mid \delta, \epsilon) \\
& = \prod_{i=1}^{D} \left\{ \prod_{j=1, j \neq i}^{D} \{ p(a_{ij} \mid s_{ij}, r_{ji}, \Psi_j) p(s_{ij} \mid \eta_i) p(r_{ji} \mid \eta_j) \} \times \right. \\
& \qquad \left. \prod_{k=1}^{K} \prod_{k'=1}^{K} p(\psi_{ikk'} \mid \delta_{kk'}, \epsilon_{kk'}) \right\}. \qquad (2.3)
\end{aligned}
$$

Figure 2.1 shows the graphical representation of the proposed model.

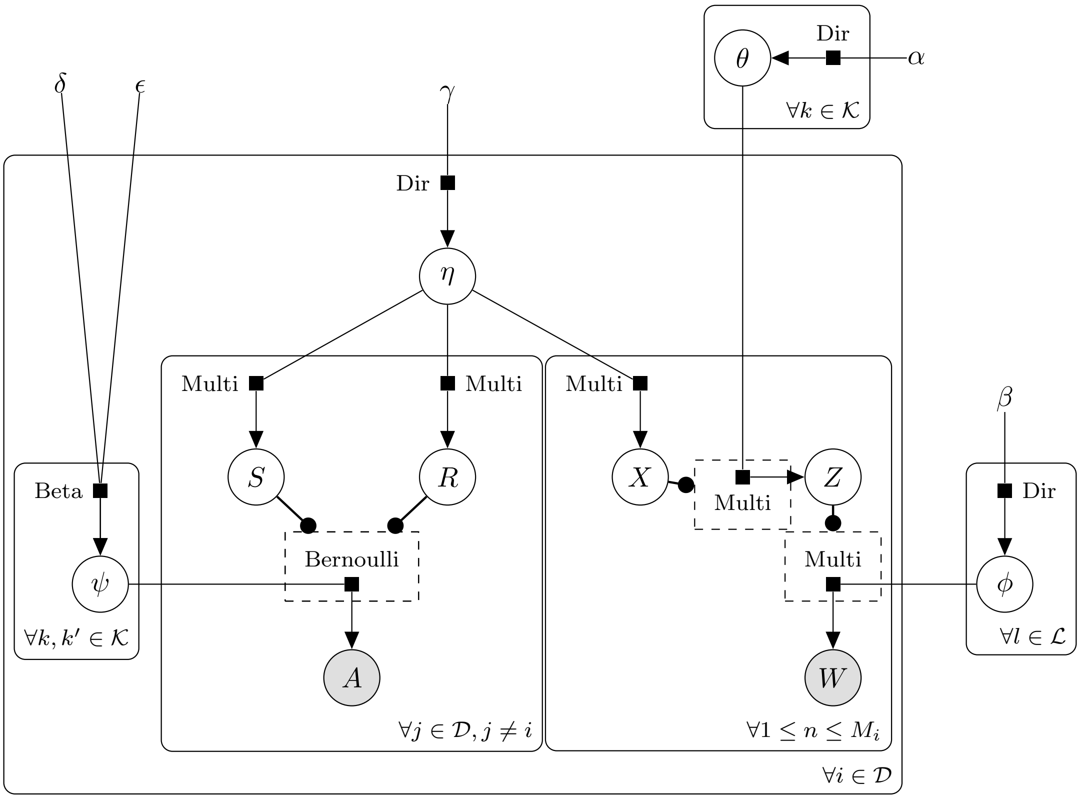

θ
Dir
α
$\forall k \in \mathcal{K}$
δ
ϵ
γ
Dir
η
Multi
Multi
Multi
β
Beta
S
R
X
Multi
Z
Dir
Bernoulli
Multi
ψ
ϕ
$\forall k, k' \in \mathcal{K}$
$\forall l \in \mathcal{L}$
A
W
$\forall j \in \mathcal{D}, j \neq i$
$\forall 1 \leq n \leq M_i$
$\forall i \in \mathcal{D}$

FIGURE 2.1: Graphical model of the proposed model

2.3.2 Estimation Procedure

In previous studies, many methods for estimating topic models have been proposed, such as variational Bayesian methods, but one of the mostly used methods is the collapsed Gibbs sampling (CGS, Griffiths and Steyvers, 2004) method. This method integrates out a part of model parameters in the process of deriving the posterior distribution, and it can conduct efficient sampling of the candidates from the posterior.

We can derive sampling equations similar to the model of Igarashi and Terui (2020) derived in Section 1.3, and by integrating four parameters, community distribution H, edge probability Ψ, topic distribution Θ, and word distribution Φ, the conditional posterior distributions of latent communities S and R are defined as follows.

$$\begin{aligned}
&p(s_{ij}=k, r_{ji}=k' \mid a_{ij}, A_{\backslash ij}, S_{\backslash ij}, R_{\backslash ji}, X, \gamma, \delta, \epsilon) \\
&\propto \int\int p(s_{ij}=k, r_{ji}=k' \mid \eta_i, \eta_j) p(x_i, x_j \mid \eta_i, \eta_j) p(\eta_i, \eta_j \mid S_{\backslash ij}, R_{\backslash ji}, X, \gamma) d\eta_i d\eta_j \\
&\quad \times \int p(a_{ij} \mid \psi_{jkk'}) p(\psi_{jkk'} \mid A_{\backslash ij}, S_{\backslash ij}, R_{\backslash ji}, \delta, \epsilon) d\psi_{jkk'} \\
&= \frac{N_{ik\backslash ij} + M_{ik} + \gamma_k}{\sum_t \left(N_{it\backslash ij} + M_{it} + \gamma_t\right)} \times \frac{N_{jk'\backslash ji} + M_{jk'} + \gamma_{k'}}{\sum_t \left(N_{jt\backslash ji} + M_{jt} + \gamma_t\right)} \times \\
&\quad \frac{\left(n^{(+)}_{jkk'\backslash ij} + \delta_{kk'}\right)^{\mathbb{I}(a_{ij}=1)} \left(n^{(-)}_{jkk'\backslash ij} + \epsilon_{kk'}\right)^{\mathbb{I}(a_{ij}=0)}}{n^{(+)}_{jkk'\backslash ij} + n^{(-)}_{jkk'\backslash ij} + \delta_{kk'} + \epsilon_{kk'}}. \qquad (2.4)
\end{aligned}$$

The conditional posterior distributions of latent word community and word topic are defined as follows.

$$\begin{aligned}
&p(x_{im}=k, z_{im}=l \mid W, S, R, X_{\backslash im}, Z_{\backslash im}, \alpha, \beta, \gamma) \\
&\propto \int p(s_i, r_i \mid \eta_i) p(x_{im}=k \mid \eta_i) p(\eta_i \mid S, R, X_{\backslash im}, \gamma) d\eta_i \times \int p(z_{im}=l \mid \theta_k) \\
&\quad p(\theta_k \mid X_{\backslash im}, Z_{\backslash, im}, \alpha) d\theta_k \times \int p(w_{im}=v \mid \phi_l) p(\phi_l \mid W_{\backslash im}, Z_{\backslash im}, \beta) d\phi_l \\
&= \frac{N_{ik} + M_{ik\backslash im} + \gamma_k}{\sum_t \left(N_{it} + M_{it\backslash im} + \gamma_t\right)} \times \frac{M_{kl\backslash im} + \alpha_l}{\sum_q \left(M_{kq\backslash im} + \alpha_q\right)} \times \frac{M_{lv\backslash im} + \beta_v}{\sum_u \left(M_{lu\backslash im} + \beta_u\right)}. \qquad (2.5)
\end{aligned}$$

In the estimation procedure using CGS, we repeatedly sample the latent communities for each relationship and latent topics for word according to the Equations (2.4) and (2.5). Finally, estimates of the four parameters integrated out in the posterior deriving process are obtained by calculating the expectations of samples that excludes burn-in samples depending on the initial values.

2.4 Empirical Analysis

2.4.1 Estimation Results of Empirical Analysis

In this section, we preset an empirical analysis using Twitter dataset to show that the analysis using the proposed model is useful for understanding real world online social networks. Since we use the same Twitter dataset as Section 1.5, we will not go into the details of dataset.

When estimating the block models including the proposed model, in general, we need to determine the number of communities (and the number of topics in this study). Previous studies have proposed various methods for determining the number of communities as a model comparison using the information criterion, such as the BIC method (Handcock, Raftery, and Tantrum, 2007; Saldaña, Yu, and Feng, 2017), the integrated completed likelihood method (Daudin, Picard, and Robin, 2008; Bouveyron, Latouche, and Zreik, 2018), and the variational Bayesian method (Latouche, Birmelé, and Ambroise, 2012). However, we adopted a widely applicable information criterion (WAIC, Watanabe, 2010), which was recently proposed as a new information criterion and is now used in many fields. The definition of WAIC for the proposed model is almost the same as the one for the model of Igarashi and Terui (2020) in Appendix B.1 so we exclude it here.

Table 2.1 shows the results of calculating WAIC for the model using Twitter dataset with the number of communities and topics ranging from 5 to 10 (K is the number of communities and L is the number of topics, and boldface indicates the lowest values in the table). The number of iterations was 5,000, of which 2,000 were excluded as the burn-in period depending on the initial value. The settings of the hyperparameters are $\alpha_l = 0.1, \forall l$, $\beta_v = 0.1, \forall v$, $\gamma_k = 1.0, \forall k$, $\delta_{kk'} = \epsilon_{kk'} = 0.1, \forall k, k'$, respectively. As a result, we selected a model with seven communities and seven topics, and we discuss the estimation results with this model in the following.

First, we look at node independent global parameters (word distributions Φ and topic distributions Θ) to see what people in the detected communities are interested in. Table 2.2 lists the top 10 words with the highest value of the estimated word distribution for each topic, which allows us to interpret the meaning of the topics. Relevant words representing each topic are underlined, and the meaning of the topics can be interpreted as follows. Topic 1: animation (such as blackov, hunterxhunt, jojosbizarreadventur), Topic 2: streaming and broadcasting (such as teamemmmmsi, twitchkitten, roku), Topic 3: music (such as vevo, spinrilla, zeldathon), Topic 5: reading books (such as amread, bookreview, kindleunlimit), Topic 6: business (such as digitalmarket, smm, contentmarket), and Topic 7: sports (such as oiler, tfc).

TABLE 2.1: Model comparison with WAIC

	$L = 5$	$L = 6$	$L = 7$	$L = 8$	$L = 9$	$L = 10$
$K = 5$	4422206.32	4340879.93	4321068.95	4333535.35	4354814.11	4553144.83
$K = 6$	4333313.32	4333488.66	4351008.38	4309479.01	4302773.27	4280703.13
$K = 7$	4313265.58	4285253.01	**4272682.48**	4346780.91	4301005.75	4414800.13
$K = 8$	4320416.87	4282485.37	4326300.05	4324393.23	4321806.29	4426226.19
$K = 9$	4429170.84	4329997.66	4439594.82	4407656.85	4296128.61	4301655.85
$K = 10$	4361219.83	4342899.53	4282056.30	4306509.44	4306244.12	4406655.34

TABLE 2.2: Top 10 words with the highest value of word distributions of the proposed model

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
nonfollow	teamemmmmsi	trapadr	criticalrol	iartg	growthhack	savvi
blackclov	dokkan	vevo	zeldathon	amread	digitalmarket	lube
hunterxhunt	twitchkitten	ddrive	orton	erotica	gdpr	foodporn
jojosbizarreadventur	vgc	leed	fursuitfriday	asmsg	smm	oiler
mkleosaga	roku	spinrilla	dramaalert	momlif	contentmarket	austria
wnf	wizebot	ifb	sdlive	hemp	gamedesign	tfc
hori	ryzen	gainwithpyewaw	htgawm	writerslif	podernfamili	crowdfir
mdva	freebiefriday	gainwithxtiandela	sml	bookreview	socialmedialmarket	tranc
hyrulesaga	streamersconnect	horford	robloxdev	kindleunlimit	bigdata	tock
nyxl	nbaliv	suav	yoongi	bookboost	emailmarket	texfil

TABLE 2.3: Top 10 words with the highest value of word distributions of LDA

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
trapadr	teamemmmmsi	vevo	nonfollow	podernfamili	growthhack	gamedesign
ddrive	tfc	spinrilla	dokkan	iartg	digitalmarket	leed
ifb	twitchkitten	htgawm	zeldathon	amread	gdpr	savvi
gainwithpyewaw	hori	beck	criticalrol	asmsg	smm	lube
gainwithxtiandela	roku	orton	vgc	erotica	contentmarket	momlif
blackclov	mkleosaga	sdlive	fursuitfriday	foodporn	socialmediamarket	quoteoftheday
hunterxhunt	wnf	horford	dramaalert	dogsoftwitt	bigdata	austria
jojosbizarreadventur	wizebot	suav	sml	oiler	cto	hemp
yoongi	ryzen	herewego	robloxdev	writerslif	emailmarket	tranc
hoseok	streamersconnect	drippin	spforstreami	amiga	fintech	tock

Figure 2.2 shows the estimated topic distribution for each community, and we can see the proportion of topics within each community. The figure shows that the topic distribution is concentrated on a single and unique topic for each community. This is probably because the structure of the proposed model is such that it extracts a set of nodes with a high density of edges and similar text topics, i.e., a topic-based community, but this cannot be distinguished from the figure alone. Therefore, we further explore the estimation results of Figure 2.2 by comparing the simultaneous approach of the proposed model, which considers both network and text information, with the independent approach, which integrates the results of two independent models: the network model considering only network information to extract the community structure and the LDA model considering only text information to extract topic structure. In the following, we compare the interpretation of topics extracted by the word distribution, the topic distribution for each community, and the estimated community structure, for the simultaneous approach and the independent approach, respectively.

First, Table 2.3 shows the relevant words on the topics extracted by LDA. The table includes the many similar words to the result of the proposed model shown in Table 2.2 in the columns of the same topics. Therefore, we can confirm that the same topics are extracted in modeling that considers both network and text and in modeling of text only.

Next, we integrate the results of the network model and the LDA model to evaluate the topic distribution for each community. While the LDA considers a document to be a word set and estimates the topic distribution for each document, here it estimates the topic distribution for each node because a word set is considered to accompany a node. Also, the network model also estimates the proportion of communities nodes belong to. Therefore, we can derive the topic distribution for each community of the independent approach, as estimated by the proposed model, by summing up the topic distributions of all nodes weighted by the community proportion. Let the topic distribution for each node estimated by the LDA model be $\hat{\lambda}_i^{(ind)}$ and the community distribution for each node estimated by the network model be $\hat{\eta}_i^{(ind)}$, and the topic distribution for each community of the independent approach

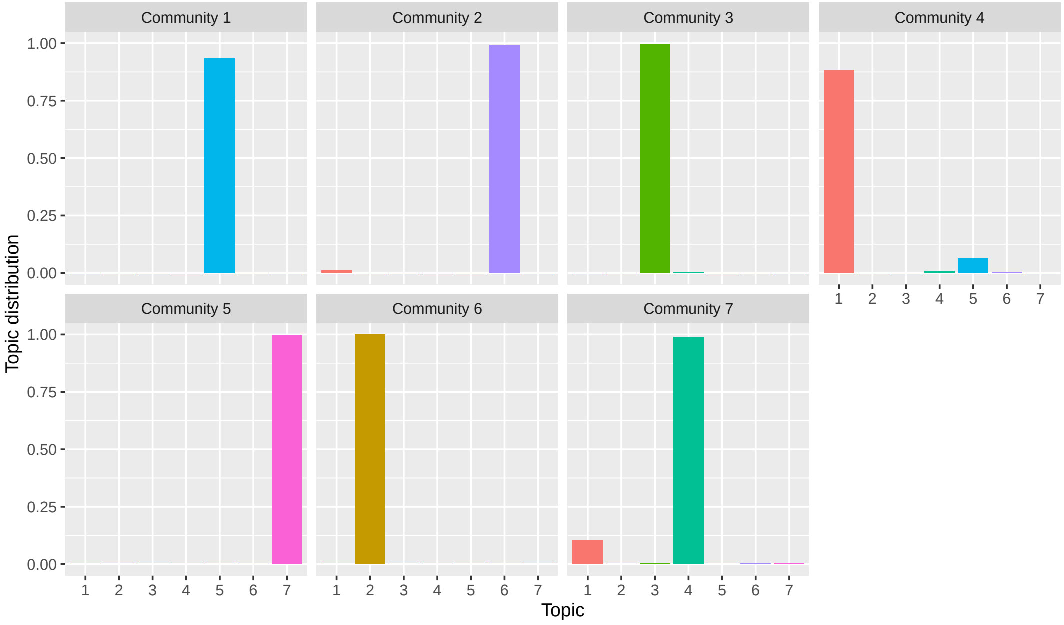


FIGURE 2.2: The estimates of topic distribution for each community of the proposed model

is derived as follows:

$$\theta_k^{(ind)} = \sum_{i=1}^{D} \hat{\lambda}_i^{(ind)} \times \hat{\eta}_{ik}^{(ind)}, \qquad k = 1, \dots, K. \tag{2.6}$$

The results are shown in Figure 2.3, which indicate that multiple topics correspond to a single community in contrast to the results of the proposed model shown in Figure 2.2.

We then compare the intra-community edge densities of the proposed model with the network model of the independent approach. Both models assume mixed membership for the network generation process; therefore, we calculate the edge density by defining the nodes' belonging community at the highest values of the estimated community distributions. Figure 2.4 shows the edge densities and the number of nodes in the community for both models (the top and bottom left figures are the results of the network only model, and the top and bottom right are the results of the proposed model). In the figure, the community numbers are reordered in order of increasing intra-community edge density (diagonal component) for comparison. The network only model found three low edge-density communities with many nodes (communities 7, 3, and 1), whereas the proposed model found two such

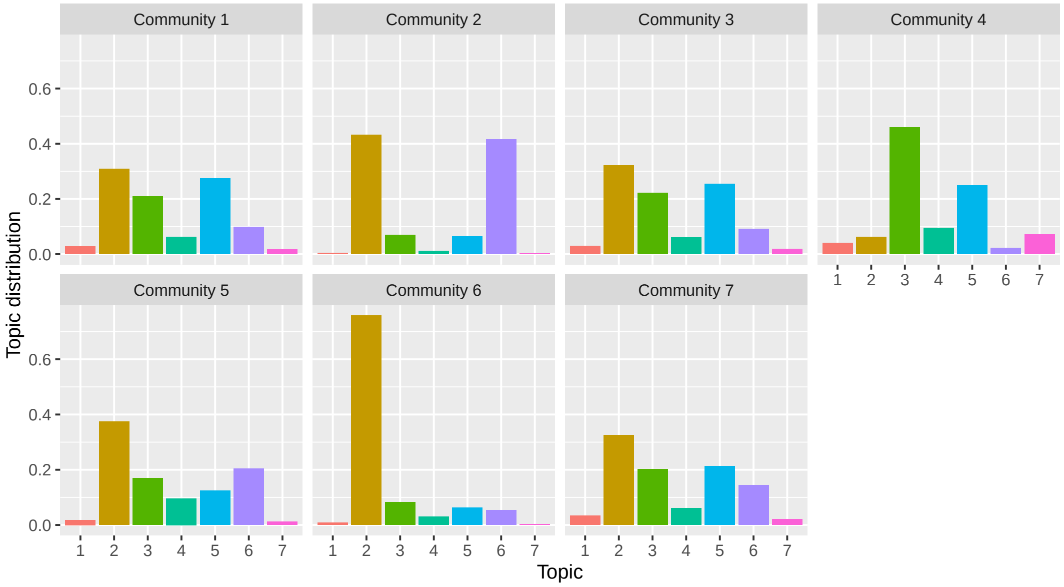


FIGURE 2.3: The estimates of topic distribution for each community of the independent approach

communities (communities 5 and 4). Thus, estimated values of the intra-community edge densities are slightly different overall between the two models. However, they both capture the same structure for the rest of the community structure. For example, they extract communities consisting of a small number of nodes (community 4 for the network only model and community 6 for the proposed model) and medium-sized communities with relatively high density of internal connections (communities 2 and 5 for the network only model and communities 1 and 3 for the proposed model).

In summary, we can say that the proposed model clearly represents the topic structure in the community while capturing the community structure and the meanings of extracted topic that are overall similar to the independent approach. However, these results are based on the dataset used in this study, and further discussions, including theoretical analysis, are needed to verify such properties in general networks.

Finally, we see the estimation results of the heterogeneous local parameters for each node (edge probability Ψ). Figure 2.5 shows the estimated edge probabilities and the community distributions for nodes 1 and 237, where the in-degree and out-degree of node 1 are 6 and 0, respectively, and those of node 237 are 657 and 37,

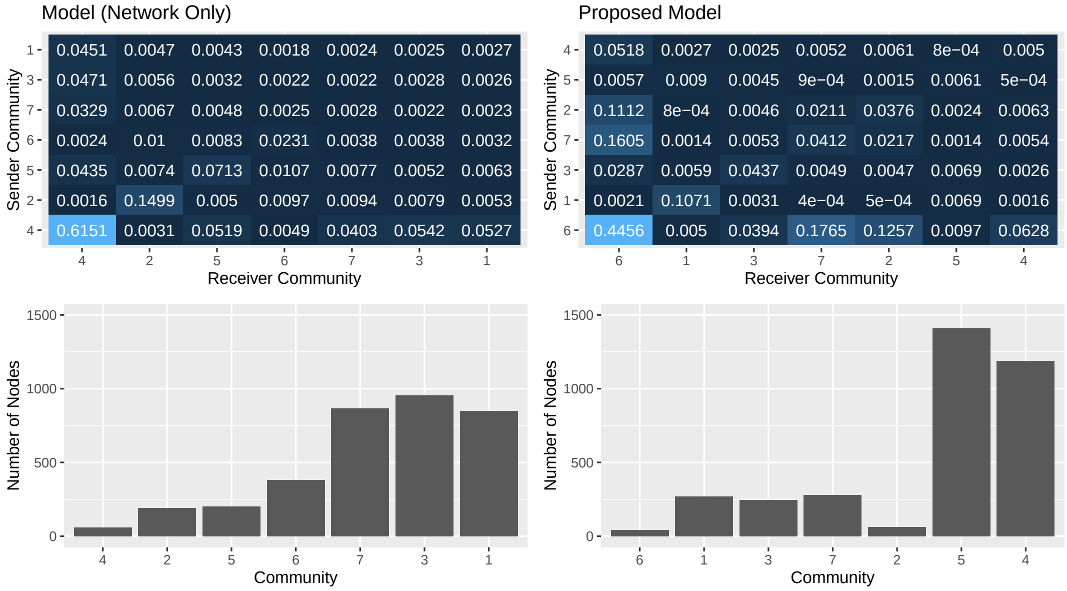
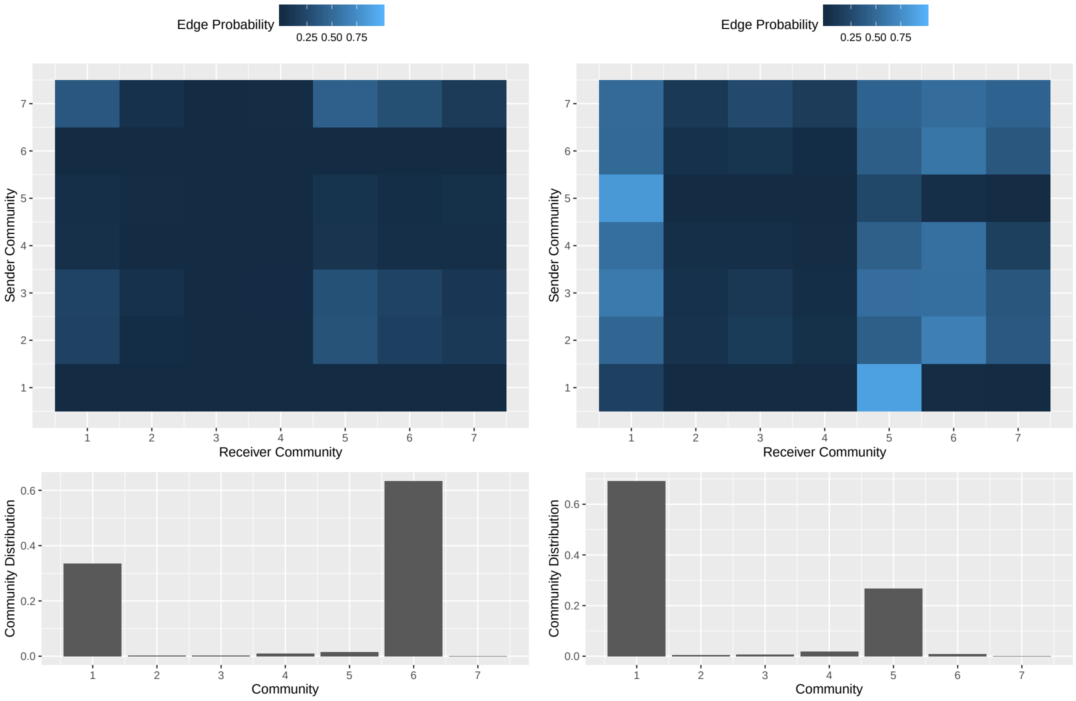


FIGURE 2.4: The edge densities and the number of nodes within the community for the proposed model and the network only model

respectively. The estimation results show the node degree heterogeneity of both nodes: the edge probabilities for node 1 are estimated at low values with respect to the communities to which node 1 mainly belongs (community 1 and 6), while them for node 237 are estimated at high values (community 1 and 5). As these results indicate, by introducing assumption that takes the node degree heterogeneity into account in the parameters of the edge probability, the model is expected to be able to represent the network mode more flexibly and improve the predictive performance on test data. In the next section, in order to verify it, we compare the proposed model with the comparative models that exclude the properties of the proposed model, consideration of the text information and the node degree heterogeneity.

2.4.2 Model Comparison of Predictive Performance

In this section, we examine the predictive performance of the proposed model for out-of-sample data and the effect of consideration of text information and the node degree heterogeneity on the predictive performance through a model comparison with the different models. While the proposed model (Model 4) includes consideration of both text information and node degree heterogeneity, we consider three comparative models: a model that considers neither text nor heterogeneity (Model



Node 1

Node 237

FIGURE 2.5: The estimated edge probability and community distribution for node 1 (left) and node 237 (right)

1), a model that does not consider text but considers heterogeneity (Model 2), and a model that considers text but does not consider heterogeneity (Model 3).

In Section 2.4.1, all network and text data were used as training data to estimate the mode, but here, 90% of the $D-1$ network data of each node were used as training data, and remaining 10% were used as test data. As for the text data, all words were used as training data as in the previous section. The number of iterations and the settings of hyperparameters are also the same conditions as in the previous section. Under these conditions, we estimate the models on the training data. Let the estimated community distribution and edge probability be $\hat{H}$ and $\hat{\Psi}$, the predictive probability of the test network data $a_{ij} \in A^{(test)}$ for the proposed model can be calculated as follows.

$$p(a_{ij}=1) = \sum_{k=1}^{K}\sum_{k'=1}^{K} \hat{\eta}_{ik}\hat{\eta}_{jk'}\hat{\psi}_{jkk'}. \tag{2.7}$$

The predictive probabilities for the rest three models can also be calculated by the product of the community distribution and the edge probability in the similar way. Here, the number of communities and topics was set to from 5 to 10, respectively, and the area under the curve (AUC) of each model was calculated in a grid. For each model and each number of communities and topics, we repeated the model estimation and the calculation of the AUC 30 times while shuffling the test data. Figure 2.6 includes box-plots summarizing the AUCs for each mode and the number of communities and topics.

First, we compare models that do not consider degree heterogeneity (Model 1 and 3) with models that do (Model 2 and 4). Without considering the degree heterogeneity, the AUC increases as the number of communities increases, which is as expected because generally, the accuracy of clustering increases with the number of clusters until an optional number is reached. On the other hand, models that consider degree heterogeneity have a higher overall AUC than homogeneous models, which indicates that the introduction of heterogeneity in the model of network generative process increases the predictive performance. However, in Model 2, the AUC declines with the increase in the number of communities, and the degree heterogeneity may have a negative aspect that worsens the predictive performance. This may indicate a disadvantage of introducing degree heterogeneity with the model specification of this study, because it reduces the data for estimating the edge probability of each node, and this effect becomes clearer as the number of communities increases. In that case, the negative effect can be improved by incorporating the common structures between nodes as factors leading the node degree heterogeneity. Such modeling that assumes common factors behind heterogeneous structures can be easily achieved through hierarchical Bayesian modeling. Also, some appropriate examples of such common factors may be node specific attributes such as gender and age, which have been studied in sociology.

Next, comparing models that do not consider text information (Model 1 and 2) with those that do (Model 3 and 4), we can see that the predictive accuracy is improved by considering the text information. Furthermore, the effect is not constant with respect to the number of communities, and as the number of communities increases, the increase in predictive accuracy due to the text effect enlarges. In other words, when dividing a community into smaller communities, text information enables more precise clustering that is robust to predict out-of-samples, even in situations where it is difficult to divide a community into smaller parts using only

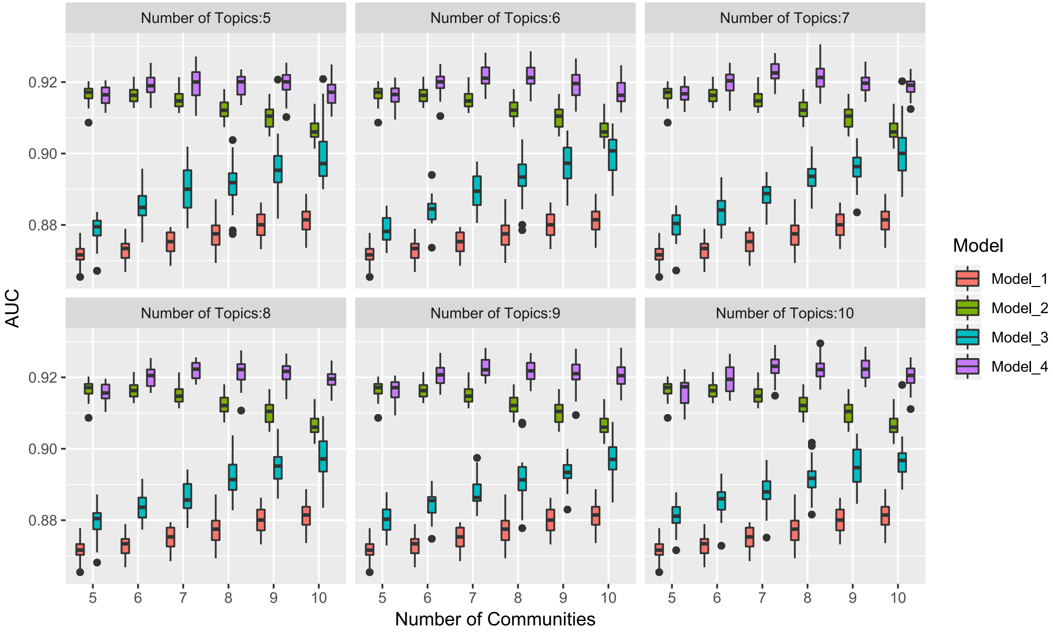


FIGURE 2.6: The values of AUC for each model and the number of communities and topics

network information. As for the proposed model (Model 4), although the degree heterogeneity have negative impacts on the predictive performance, it and another positive effect due to the text information counterbalances each other, so that the predictive accuracy does not decrease as the number of communities increases and maintains the highest AUC among the comparative models.

As a result, this model comparison shows that the proposed model that considers both text information and node degree heterogeneity has the best predictive performance compared with the different models that do not consider the respective features. However, some issues are still unresolved: the proposed model specification with respect to the degree heterogeneity may worsen the predictive performance, which is left for future works, including model improvement.

2.5 Conclusion

In this study, we proposed a model that takes into account not only network information but also text information on social media, which expresses people's interests, to make social network analysis more meaningful and realistic. Furthermore, we extended the existing network model by allowing the edge generation probability to vary for each node to consider node degree heterogeneity, which is a remarkable characteristics in real social network. This extension allows us to detect topic-based communities that have the structure of dense edge clusters and node specific text data generated from the same distribution, with sufficient goodness of fit for general social networks where the node degree varies widely among nodes.

The empirical analysis shows that the proposed model estimated by collapsed Gibbs sampling not only captures interpretable community and topic structure for real-world Twitter data, but also has better predictive performance than comparative models that do not consider text information or degree heterogeneity. Furthermore, we found that taking into account the text information of each node is useful to improve the prediction of out-of-samples when analyzing clusters in more detail, rather than taking a broad view of the community structure of the network.

Since this study focused on online social network, it was assumed that people consider only the network and text information when they connect with other people on the network. Hence, we do not consider other node-specific features, some attributes such as age and gender, or information on people's behaviors and attitudes because of data availability. In the literature on social network analysis, a number of studies have shown that such node-specific (or diad and triad) features influence network formation (Hoff, Raftery, and Handcock, 2002; Handcock, Raftery, and Tantrum, 2007). In this study, the function of the edge generation consists of the community distribution and the edge probability of receiver node, but this function can be meaningfully extended to include node-specific features such as attributes and behavioral information. We would like to incorporate this information into the model and the data availability in future work.

Chapter 3

A Dynamic Topic Model for Social Influence of User Generated Contents on Social Media

3.1 Introduction

In modern social media, a lot of people take various behaviors on social media (generating contents, connecting with others, rebroadcasting, etc.). These behaviors sometimes affect behaviors of others who observe their behaviors. This effect is called social influence, which is defined as the effects of actions taken by individuals on others' actions, attitudes, or emotions on their social network.

In the fields of marketing and sociology, many researchers have studied on social influences from various perspectives, such as word-of-mouth effects (Godes and Mayzlin, 2004; Chevalier and Mayzlin, 2006) and relationships with social network structure (Granovetter, 1973; Iyengar, Van den Bulte, and Valente, 2011). In this study, we focus on topics of social influence, which has gradually gained attention. For example, consider an individual who likes music and sports. He posts and rebroadcasts some contents related to music and sports via own social media account like Twitter or Facebook. If his behaviors have significant social influence, peers connecting with him on the social media will be affected by his behaviors in some way. However, do his generating contents of music and sports equally have the same extent of social influence? While his recommendations of music artists may positively change peers' attitudes toward the artists, his deliberations about favorite baseball team may have no effect on peers. This aspect of social influence is often discussed in the context of effects of adoption behaviors of products varying on product characteristics, such as hedonic, functional, and utilitarian products (Schulze, Schöler, and Skiera, 2014; Park et al., 2018). Meanwhile, this study focuses on social media users' behaviors, especially user-generated-contents, UGC. That is, the purpose of this study is to capture social influence that are heterogeneous across not only users but also topics of UGC.

Understanding user- and topic-specific influence plays a critical role in the current digital marketing environment. For example, it helps us make target marketing more effective. In delivering advertisements, firms can take into account not only the strength of influence consumers have but also the topic they have greater influence. Current target marketing technology considers consumers' overall social influence (e.g., Trusov, Bodapati, and Bucklin, 2010), but not the influence of each topic. Also, some media streaming services (e.g., Apple Music) recommend their contents which users have already adopted to other users. By considering the user- and topic-specific influences, the platform managers can effectively arrange the contents, and it improves the recommendation systems because they often face the challenge of optimizing the limited space for the recommendation.

The rest of this chapter is organized as follows: related literature is discussed from the theoretical perspective in Section 3.2 and the methodological perspective in Section 3.3. Section 3.4 introduces the proposed dynamic topic model for estimating social influences and derives the estimation procedure. The empirical analysis using a real social media dataset of Section 3.5 demonstrates the effectiveness of the proposed model. Finally, Section 3.6 provides some concluding remarks.

3.2 Literature Review from the Theoretical Perspectives

Social influence has been studied well from various perspectives, especially in the field of marketing and sociology. In this section, we organize them from the following three perspectives to clarify the positioning of this study within the tons of literature: "What kinds of behaviors have social influence?" "How do we measure the behaviors?" "Where does social influence occur?"

3.2.1 What Kinds of Behaviors Have Social Influence?

In the literature, researchers have focused on many kinds of behaviors to confirm the existence of social influence, and consumer behaviors of product reviews, or word-of-mouth (WOM), received great attention. They primarily consist of product ratings and review texts, and a number of studies have investigated the former's influence in the several industry of television (Godes and Mayzlin, 2004), movies (Liu, 2006; Chintagunta, Gopinath, and Venkataraman, 2010), books (Chevalier and Mayzlin, 2006), and video games (Zhu and Zhang, 2010). Also, more recent attention has focused on the latter, for example, Sonnier, Mcalister, and Rutz (2011) look beyond ratings and investigate the sales effect of the volume of positive, negative, and neutral reviews using automated sentiment analysis.

In addition to product review, consumers may be stimulated to purchase products by observing others adoption behaviors. Traditionally, the role of cumulative number of consumers on current new adoption is well studied (e.g., Bass, 1969), whereas in recent years, there has been an increasing amount of literature on the influence of individual adoptions, especially adoptions of influentials (Nair, Manchanda, and Bhatia, 2010; Wang, Aribarg, and Atchadé, 2013). Also, historically, the two streams of WOM and observational adoption have been separately discussed, but recent several studies simultaneously explore their effects (Chen, Wang, and Xie, 2011; Ameri, Honka, and Xie, 2019).

Besides such behaviors, researchers have studied on the roles of various behaviors, for example, user generated videos (Yoganarasimhan, 2012) and music (Lanz et al., 2019), rebroadcasting (Gong et al., 2017), logging web sites (Trusov, Bodapati, and Bucklin, 2010), and shopper behaviors (Zhang et al., 2018).

3.2.2 How Do We Measure Behaviors?

In the study on social influence, the types of focal behaviors and how these behaviors are measured are critical. A fundamental measure is the volume of behavior is one of the most fundamental measures, for example, the number of posted product reviews (Godes and Mayzlin, 2004), the number of peers who adopted the product (Bollinger and Gillingham, 2012), and the length of review texts (Lu, Wu, and

Tseng, 2018). Another measure, apart from volume, is the valence of behavior. It is used as a measure representing positivity or negativity of the focal behaviors, such as the average of product review ratings (Moe, Trusov, and Smith, 2011) and the emotions of review texts (Sonnier, Mcalister, and Rutz, 2011; Wu et al., 2015). Also, Moe, Trusov, and Smith (2011) employed not only the volume and valence but also variance of WOM, which is measured as the variance of ratings, and found that high variance, that is, disagreement among those who post reviews, tends to discourage the posting of extreme opinions by subsequent posters.

The existing literature often use these measures of volume, valence, and variance as primal ones. In addition to them, another measure that has recently attracted attention is behavioral content, which is also the focus of this study. Even if the same pairs of agents are on the same social network, their social influence may differ according to their adopted products and created contents. Some studies examine the difference in influence across adopted products. For example, Wang, Aribarg, and Atchadé (2013) show that while expert individuals exert greater influence on technology-related products (e.g., Bluetooth headset), which are dominant by informational influence, popular individuals exert greater influence on fashion-related products (e.g., sports paraphernalia), which are dominant by expressive-value influence. Also, Schulze, Schöler, and Skiera (2014) focus on the difference between high and low utilitarian products and Park et al. (2018) focus on the difference between hedonic and functional products, and both of them report that the characteristics of products make the extent of social influence differ.

However, in these studies, the product characteristics must be antecedently determined, and the difference in social influence is examined only across the predetermined dimensions. These are critical limitations for this study because we aim to find specific individuals with disproportional influence in the situation where several users generate contents with various topics. Determining the characteristics of such contents in advance is difficult, and discussing the differences only across the predetermined dimensions is inappropriate. For this limitation, as will be discussed later, this study uses topic modeling to estimate heterogeneous influence and its dimensions simultaneously.

3.2.3 Where Does Social Influence Occur?

Apart from the types and measures of behavior, the literature on social influence has highlighted the importance of situations where social influence occurs, that is, the structure of the social network. Many researchers have studied on the effect of global characteristics of the focal network, such as degree distribution (Dover, Goldenberg, and Shapira, 2012) and cluster coefficient (Choi, Kim, and Lee, 2010). Also, as individual characteristics on the network, some centrality measures are often applied as proxies for the position of those who have influenced or been affected (Cho, Wang, and Lee, 2012; Susarla, Oh, and Tan, 2012).

In addition, the relationship between two (or more) individuals is also important factor. Granovetter (1973) proposes that the relationships are characterized as tie strength, which is conceptualized as "a combination of the amount of times, the emotional intensity, the intimacy, and the reciprocal services which characterize the tie" as Granovetter (1973, pp.1361). Starting with his proposal, a large body of literature has investigated the impact of tie strength. Although his work and subsequent works following it claim that weak ties have a greater impact than strong ties, some different findings are conflicting them (refer to Muller and Peres, 2019, for a comprehensive survey on tie strength). Also, (Zhang and Godes, 2018) investigated the dynamic effect of tie strength on decision quality and reveal that at the beginning of community participation, information from strong ties is reliable in improving the decision-making quality. However, once one has sufficient experience in the community, the effect of weak ties increases and becomes consistent with the stream of Granovetter (1973). In marketing, not equivalent, strong and weak ties are often replaced into personal and community networks (Ameri, Honka, and Xie, 2019). On the one hand, personal network, suchh as Twitter and Facebook, is relationships built by following acquaintances, on the other hand, community network, such as Amazon and Yelp, is relationships built by being affected by strangers' behaviors, such as product reviews and adoptions.

While these studies target the influence on a single network, other studies expand the discussion into multiple networks. Iyengar, Van den Bulte, and Valente (2011) analyze the influence on doctors' prescription behaviors using two networks, "discussion" and "referral." Similarly, Chen, Van der Lans, and Phan (2017) use four relationships, "economic," "social," "religious," and "family," to study the influence on adoptions of micro-finance program.

3.2.4 Positioning and Contribution of This Study

In the above discussion, we organized a large and growing body of literature having investigated social influences from various perspectives into the three streams. Here, we clarify the positioning of this study in the literature according to these streams: we investigate social influence of UGC measured by topics in a single personal network (strong ties) on social media. To achieve this goal, as discussed later, we propose a topic modeling to capture such heterogeneout influences across individuals and topics of UGC. All points specified in the literature are difficult to incorporate into a single research; hence, examining other interesting combinations (e.g., topic-specific influence on weak ties) is beyond the scope of this study.

This study contributes to existing social influence research in two ways. First, we identify differences in social influence across topics of UGCs on social media. Existing studies focus on the influence of adoption of products (Wang, Aribarg, and Atchadé, 2013; Schulze, Schöler, and Skiera, 2014; Park et al., 2018), and to our best knowledge, this is the first study to explore the topic-specific influence on UGC of social media. Second, we simultaneously estimate the topics of UGC and social influence varying for the topics. In previous studies, researchers must know the characteristics of products or behaviors beforehand (e.g., hedonic and functional products). However, in a social media environment, the dimensions of UGC with various topics are hard to determine using only prior knowledge. The use of topic modeling can deal with this limitation.

3.3 Literature Review from the Methodological Perspectives

In this section, we discuss some approaches of literature for three issues related to estimation of social influence, social spillover and multiplier, identification problems, and topic modeling.

3.3.1 Social Spillover and Social Multiplier

Social interactions are of key interest to marketers and policy makers due to the presence of social spillovers and multipliers. A social spillover arises when an intervention or marketing action to an agent affects the behavior of others in the agent's reference group via a social interaction. In some cases, the agent may have influence on others and may be influenced by others. In such case, social interactions engender a social multiplier that multiplies the effect of the intervention to the initial agent by the feedback loop. Hartmann et al. (2008) characterize such relationships with social multipliers as active interactions in contrast to passive interactions where only social spillovers arise.

Traditionally, spillovers have been estimated by the Bass model (Bass, 1969). But, while it includes the structure to estimate spillovers that current adoptions of agents are affected by the cumulative adoptions of others who have adopted, the others are assumed not to be affected by the agents. Hence, the Bass model considers only passive interactions. Some previous studies use spatial models for capturing spillovers (Toker-Yildiz et al., 2017; Zhang, 2019). Given the spatial weight matrix corresponding to social relationships, spatial models capture the interdependence of agent behaviors. In addition, this model can consider spillovers by using asymmetric spatial matrix, multipliers by using symmetric spatial matrix, and their heterogeneity by using a weighted spatial matrix. These spatial matrices, however, are observed data, hence, this model cannot estimate individual social influence.

Although no study has been conducted on estimating heterogeneous social influence using above models, some studies use hierarchical Bayes (Narayan, Rao, and Saunders, 2011) and approximate Bayes (Trusov, Bodapati, and Bucklin, 2010) to allow parameters to vary for individuals and estimate heterogeneous social influence. Narayan, Rao, and Saunders (2011) model that consumers update their preferences when being affected by self-reported influencers having heterogeneous influences. Also, Trusov, Bodapati, and Bucklin (2010) propose a model for capturing social media users' login behaviors affected by others' login and demonstrate that the proposed model can estimate heterogeneous influence for each tie to detect influencers.

A simple model for multipliers is linear-in-means model. This model assumes that agents' behaviors are linearly correlated to means of others' behaviors in the reference groups of the agents. Nair, Manchanda, and Bhatia (2010) use linear-in means model to reveal that doctors' prescribing behaviors are affected by the average prescription of other doctors in the same district and influencer's prescription. Some studies estimate multipliers using a linear model, apart from linear-in means; for example, Nam, Manchanda, and Pradeep (2010) find that adoptions of video-on-demand have linear effects on peers' adoption behaviors. Also, Haenlein (2013) estimate peer influence of churn decisions by linear polynomial regression model. What these models have in common is that they show that all agents are equally influenced from peers in the reference group, that is, multipliers, by assuming homogeneity in the coefficients corresponding to the (often aggregated) peers' behaviors. By contrast, the above heterogeneous models can be more useful because they estimate spillover for each individual tie and consequently, they capture multipliers for some ties where bidirected spillovers exist.

Another approach for multipliers is agent-based modeling (Libai, Muller, and Peres, 2013; Peres and Van den Bulte, 2014; Phan and Godes, 2018). Libai, Muller, and Peres (2013) conduct agent-based modeling approach for evaluating WOM seeding program. Also, a recent simulation study by Phan and Godes (2018) indicates that, in contrast to earlier findings, imitators who collect information from multiple sources in the peer network have greater influences than independents that can access information outside the network. These approaches, however, simulate agents' behaviors under some assumptions, and hence, although they are sufficient for theorization, they do not estimate individual social influence on social media users' activities, which is the purpose of this study.

The proposed model in this study is based on the established framework by Trusov, Bodapati, and Bucklin (2010), and it estimates heterogeneous spillovers in the situation that individuals' contents generations are affected by peers' UGC on personal network. The benefit of this model is that we can detect influencers for each topic of UGC, and these topic-specific influence on UGC are not yet clear in the literature. Other approaches for estimating influencers' impacts are the use of questionnaires to observe influencers. There are two basic approaches currently being adopted, self-reported method (e.g., Wang, Aribarg, and Atchadé, 2013), asking survey respondents to report to what extent they perceive themselves to be influential, and sociometric method (e.g., Nair, Manchanda, and Bhatia, 2010), wherein each survey respondent reports that other respondents' opinions are incorporated in his/her own decisions. Also, some studies show that these two methods observe different perspectives of influencers (Iyengar, Van den Bulte, and Valente, 2011; Risselada, Verhoef, and Bijmolt, 2016; Zhang et al., 2018). However, although the effectiveness of these methods has been demonstrated in the literature, applying them for large-scale networks on social media is infeasible. Therefore, we take the stand that social influence should be estimated from actual behavioral data.

3.3.2 Identification Problem

The literature of social influence has recognized some issues that we should consider for identification. First, on a general social network, individuals tend to connect with those who are similar to them, known as endogenous group formation or homophily. Existing homophily implies that there is positively correlation between behaviors of individual and peers, hence, we should distinguish peer influences from correlation because of homophily. The traditional popular approach for avoiding this problem is introducing fixed effects (Nair, Manchanda, and Bhatia, 2010) or heterogeneous parameters (Hartmann, 2010). Also, Bhattacharya et al. (2019) distinguished influences from homophily by coevolution model of ties creation and contents generations. In this study, we deal with this problem by assuming heterogeneity into model parameters corresponding to peer influence.

Second, if the error term contains unobserved variables correlating both independent and dependent variables (such as common advertising to some agents and seasonality trend), estimates of coefficients may seriously have bias. For this problem, it is effective to use propensity score matching (Aral, Muchnik, and Sundararajan, 2009) or introduce fixed effects (Nair, Manchanda, and Bhatia, 2010). Hence, this study also introduces terms of agent-specific and time-specific fixed effects to mitigate the correlated unobservable issue.

The final issue is simultaneity, or reflection problem (Manski, 1993), which is that agents are affected by others in their reference group, and at the same time, the others are affected by the agents. Simultaneity causes biased estimation of coefficients because of dependence between explanatory variables and error term. Previous studies apply instrument variables (Nair, Manchanda, and Bhatia, 2010) and modeling of equilibrium (Hartmann, 2010) for eliminating effects of simultaneity. Also, some studies concern this problem by effectively designing how to observe behavioral data, for example, Wang, Aribarg, and Atchadé (2013) conduct field experiments to observe only one behavioral data for each time interval. We can avoid this problem if the data are collected at the frequency that corresponds to the true decision interval, but simultaneity will disappear when the data are frequently sampled (Franses, 2005). However, this study follows Trusov, Bodapati, and Bucklin (2010) and assumes that the sparsity for peer influences will mitigate simultaneity issues. Bidirectional peer influences cause biased estimation of coefficients because of simultaneity, and thus, introducing sparsity may decrease the probability of bidirectional peer influences, thereby leading to alleviating simultaneity issue. In addition, this sparsity assumption represents that while a few consumers have higher influences, many others have relatively lower influences, which is consistent with the findings of literature.

3.3.3 Topic Modeling

This study uses topic modeling for not quantitatively but qualitatively consider UGC on social media. Originally topic modeling, such as latent Dirichlet allocation (Blei, Ng, and Jordan, 2003, LDA,), has been developed in the field of natural language process, and they model the generative process of words using latent variables (i.e., topics).

In recent years, marketing literature has also applied topic modeling for qualitatively incorporating text data such as product reviews into marketing models. The earliest study by Tirunillai and Tellis (2014) apply topic modeling for online product reviews to map and segment brands into the dimension of the extracted topics. Also, Büschken and Allenby (2016) propose the extended topic model considering sentences of product reviews to describe how consumers use their experiences to evaluate products.

Some studies incorporate various text data, apart from product reviews, by topic modeling. For example, Zhang, Moe, and Schweidel (2017) applied LDA for postings of Twitter to create feature vectors of the postings represented by the dimension of topics. They investigate the effect of similarity between the feature vector of viewed postings and mean vector of viewer's postings on their rebroadcasting behaviors. Also, a recent work of Toubia et al. (2019) reflects the findings of literature into text analysis, in contrast to above-mentioned works automatically extracting topics from text data. They propose guided LDA extracting product features from descriptions of movies according to psychology themes for entertainment products based on media psychology.

What these models have in common is that they translate unstructured text data into qualitative feature vectors with dimensions of topics. In the same way, our proposed topic model also translates UGC into user-specific proportional vectors, and these vectors represent topic proportions in UGC for each user. We aim to estimate user- and topic-specific social influence by capturing temporal changes of these proportional vectors when the users observe peers generated contents, that is, the users generated contents are affected by the peers' contents. Certainly, other machine learning methods can be applied for extracting features from text data (e.g., word2vec and VAE), but a major advantage of topic modeling is that it can be treated as descriptive model having interpretable structures, and therefore, the key assumption that peers' behaviors have influence on the user's behaviors can be structurally incorporated into the model in the manner of Bayesian modeling.

3.4 Model

This section describes the proposed dynamic topic model for estimating social influence for each topic of user generated contents. We observe UGC on social media, and to fit the topic model for the UGC data, the contents are assumed to be decomposed into multisets of the smallest unit constructs (e.g., words in a sentence or a document and objects in a photo content) as far as can be interpreted for the meaning. In the following, we denote the UGC data as $W = \{w_{dt}\},\ d = 1, \ldots, D,\ t = 1, \ldots, T$, where

D and T is the number of users and the time period. w_{dt} consists of a multiset of constructs of contents posted by user d at time t, but if the user did not post anything at that time, w_{dt} denote an empty set.

The current study estimates social influence on social media, therefore, we also observe the follow relationship network among users, and the set of friends that user d follows is represented by F_d. For simplicity, we assume that the network does not change during the observation period of the data, hence F_d's subscript does not have a time indicator t. In practice, we may see behaviors such as unfollowing friends or following new users during the observation period. However, the bias from assuming the network as static can be reduced by taking some appropriate filtering process, such as exempting new users who have registered because they tend to change their networks frequently.

3.4.1 Model Specification

Below, we define the dynamic topic model for social influence of UGC on social media, which combines the dynamic topic model (DTM, Blei and Lafferty, 2006) for capturing content generating behaviors and the hierarchical structure of topic proportion for each user influenced by the friends' generated contents, that is, social influence, by using vector autoregression (VAR) model.

The observed UGC data $w_{dt} = (w_{dt1}, \ldots, w_{dtN_{dt}})^\top$ is a multiset of UGC elements (e.g., words and objects) posted by user d at time t, and N_{dt} denotes the number of elements within the postings by the user d at t. The topic assignment for the n-th element is assumed to follow a categorical distribution, $z_{dtn} \sim categorical(\theta_{dt})$, where $\theta_{dt} = (\theta_{dt1}, \ldots, \theta_{dtK})^\top$ (K is the number of topics) represents the topic proportion within the contents posted by user d at time t. When the latent topic assignment z_{dtn} is given, the corresponding element $w_{dtn} \in \{1, \ldots, V\}$ is also assumed to follow a categorical distribution, $w_{dtn} \mid z_{dtn} = k \sim categorical(\phi_k)$, where $\phi_k = (\phi_{k1}, \ldots, \phi_{kV})^\top$ (V is the number of unique elements) is the element distribution representing element generation probability. The matrix representation of the element distribution is denoted by $\Phi = (\phi_1, \ldots, \phi_K)$, and each element distribution is assumed to follow a Dirichlet distribution as prior, $\phi_k \sim Dirichlet(\phi_0)$, where ϕ_0 is a V dimensional hyperparameter vector.

Next, we introduce the hierarchical structure of topic proportions in which the topic distribution evolves, influenced by the friends' generated contents. We consider linear models for the hierarchical structure for the topic distributions, thus, we assume the SoftMax transformation of the topic distributions and the Gaussian linear VAR model for their prior distributions as follows.

$$\theta_{dtk} = \frac{\exp(\eta_{dtk})}{\sum_{k'} \exp(\eta_{dtk'})}, \quad k = 1, \ldots, K \tag{3.1}$$

$$\eta_{dt} = \begin{cases} \eta_{dt-1} & (w_{dt} = \emptyset) \\ \alpha_d \odot \eta_{dt-1} + \sum_{f \in F_d} \beta_{df} \odot \bar{\eta}_{ft-1} + \gamma_t + \epsilon_{dt} & (otherwise), \end{cases} \tag{3.2}$$

where $\odot$ represents the element product, and we set $\eta_{dtK} = 0$ for model identifiability.

The first term of Equation (3.2) shows self-influences by the last own topic distribution. The second term is the sum of social influence by the topic distribution of user d's friends. However, not all users post at every time point, so some friends have posted multiple times since the user d last posted, while others have not posted at all. Therefore, the average topic distribution $\bar{\eta}_{ft-1}$ reflects the friends' posting behaviors in the period from the last posting by the user d (let that time be s) to the last time from the current time $(t-1)$ as follows.

$$\bar{\eta}_{ft-1} = \begin{cases} 0 & (w_{fr} = \emptyset) \quad \forall r \in \{s, s+1, \ldots, t-1\} \\ \frac{1}{t-1-s} \sum_{r=s}^{t-1} \eta_{fr} & (otherwise). \end{cases} \tag{3.3}$$

Therefore, the coefficients of the friend average topic distribution β_{df} capture the social influence of the friend f's generated contents on the user d's content generating behaviors. Prior structure for β_{df} will be introduced in the next section.

As for the rest term, γ_t is time-specific random effect for capturing the time trend of the topic proportion. The last term ϵ_{dt} is the error term which is assumed to follow the zero-mean and identity covariance multivariate normal distribution $\epsilon_{dt} \sim MVN(0, \Sigma), \Sigma = I$. Also, the initial vectors of topic proportion independently follow the normal distribution, $\eta_{d0k} \sim N(\mu_0, \sigma^2_{\eta 0})$, $k = 1, \ldots, K-1$.

3.4.2 Shrinkage Prior for the Social Influence Coefficient

In this section, we introduce the shrinkage prior distribution for the social influence coefficient parameter defined in the previous section. The reason why we set shrinkage priors for the parameters of social influence is mainly the heterogeneity of influence on the social networks. Some studies on social network models formulate their models with the assumption of node degree heterogeneity (e.g., Karrer and Newman, 2011), which is a characteristic that few nodes have many friends but other many nodes have a few connections. We can say the same thing in the context of social influence because it is expected that the average users on the network track just a few friends, hence most of the influence β_{df} is likely to be zero. In the literature, for example, Trusov, Bodapati, and Bucklin (2010) assume latent binary variables representing whether or not the user distinctively influences on the friends behind continuous parameter of social influence. This also means the Bayesian shrinkage prior.

In our setting, on the other hand, we apply a Dirichlet-Laplace prior distribution proposed by Bhattacharya et al. (2015) for social influence parameters to ensure the sparsity in the distribution of social influence and the simple estimation procedure. This study assumes that most of social influences take values close to zero, regardless of the user pairs or topics, while only a few parameters take values away from zero. Hence, let $\tilde{\beta} = (\beta_{111}, \ldots, \beta_{DF_dK})^\top$ be the stacked vector of social influence, and then each element $\tilde{\beta}_j$, ($j = 1, \ldots, J$, J is $K \times \sum_d F_d$) follows a Dirichlet-Laplace prior as follows

$$\tilde{\beta}_j \sim Laplace\left(\delta_j \tau\right), \quad \delta_j \sim Dirichlet\left(\frac{1}{J}, \ldots, \frac{1}{J}\right), \quad \tau \sim gamma\left(1, \frac{1}{2}\right). \tag{3.4}$$

Furthermore, by the data augmentation of auxiliary variable following exponential distribution, $\xi_j \sim exp(\frac{1}{2})$, the above prior distribution of $\tilde{\beta}_j$ can be rewritten as the normal distribution, $\tilde{\beta}_j \sim N(0, \xi_j \delta_j^2 \tau^2)$. Therefore, given the topic distributions, we can easily estimate these social influence parameters including hyperparameters through the straightforward Gibbs sampling for the coefficients of the normal linear regression.

Then, the joint likelihood of the proposed dynamic topic model is given as follows.

$$
\begin{aligned}
& p(W, Z, H, \Phi, \alpha, \beta, \gamma, \Sigma, \phi_0, \xi, \delta, \tau) \\
&= p(W \mid Z, \Phi) p(Z \mid H) p(H \mid \alpha, \beta, \gamma, \Sigma) p(\phi \mid \phi_0) p(\beta \mid \xi, \delta, \tau) p(\alpha, \gamma, \Sigma) \\
&= \prod_{d=1}^{D} \prod_{t=1}^{T} \left\{ \prod_{n=1}^{N_{dt}} p(w_{dtn} \mid z_{dtn}, \Phi) p(z_{dtn} \mid \eta_{dt}) \right\} p(\eta_{dt} \mid \eta_{\cdot t-1}, \alpha_d, \beta_d, \gamma_t, \Sigma) \times \\
& \quad p(\Phi \mid \phi_0) \prod_{j=1}^{J} p(\tilde{\beta}_j \mid \xi_j, \delta_j, \tau) p(\alpha) p(\gamma),
\end{aligned}
\tag{3.5}
$$

where $p(\alpha)$ and $p(\gamma)$ are prior distributions, which are defined in the Appendix A.2.

3.4.3 Estimation Procedure

This section describes the estimation procedure for the proposed model. However, we can easily derive the estimation equations of the collapsed Gibbs sampling (Griffiths and Steyvers, 2004) for the latent topic assignment variables and the Gibbs sampling and the Metropolis-Hastings sampling for the coefficient parameters of the topic linear regression in Equation (3.2), and these are described in detail in the Appendix A.2.

Then, we derive the estimation procedure for the topic distributions. The topic distributions cannot be represented as a closed form of the posterior distribution in the straightforward way because they have the normal prior and the likelihood of categorical distribution through the SoftMax transformation. In this study, we transform the categorical likelihood into normal kernels by the Pólya-Gamma data augmentation (Polson, Scott, and Windle, 2013; Glynn et al., 2019), and then, we construct a Gibbs sampler considering the time dependency by the forward filtering and backward sampling (FFBS, Cater and Kohn, 1994) method.

From the model specification, we have $\eta_{dt} \mid \eta_{dt-1}, \{\eta_{ft-1}\} \sim MVN(\mu_{dt}, \sigma_\epsilon^2 I)$, where $\mu_{dt} = \alpha_d \odot \eta_{dt-1} + \sum_{f \in F_d} \beta_{df} \odot \bar{\eta}_{ft-1} + \gamma_t$, $\sigma_\epsilon^2 = 1$, hence $\eta_{dtk} \perp \eta_{dtk'} (k \neq k')$, and we see only the kth topic in the following.

The initial value of the topic distribution independently follows the normal distribution $\eta_{d0k} \sim N(\mu_{0k}, \sigma_{\eta 0}^2)$, thus, $\eta_{d1k} \sim N(\pi_{d1k}, \rho_{d1k}^2)$, where $\pi_{d1k} = \alpha_{dk}\mu_{0k} + \sum_{f \in F_d} \beta_{dfk}\mu_{0k} + \gamma_{1k}$, $\rho_{d1k}^2 = \sigma_{\eta 0}^2(\alpha_{dk}^2 + \sum_{f \in F_d} \beta_{dfk}^2) + \sigma_\epsilon^2$. This corresponds to the prior structure of the topic distribution. Next, the likelihood of the topic distribution for the topic assignments is defined as follows.

$$\begin{aligned} p(z_{dt} \mid \eta_{dt}) &\propto \left(\frac{\exp(\eta_{dt1})}{\sum_{k'} \exp(\eta_{dtk'})}\right)^{N_{dt1}} \cdots \left(\frac{\exp(\eta_{dtK})}{\sum_{k'} \exp(\eta_{dtk'})}\right)^{N_{dtK}} \\ &\propto \left(\frac{1}{1+\exp(\psi_{dtk})}\right)^{N_{dt}-N_{dtk}} \left(\frac{\exp(\psi_{dtk})}{1+\exp(\psi_{dtk})}\right)^{N_{dtk}} \\ &\propto \exp\left(\kappa_{dtk}\psi_{dtk} - \frac{\zeta_{dtk}}{2}\psi_{dtk}^2\right), \end{aligned} \tag{3.6}$$

where $\psi_{dtk} = \eta_{dtk} - \log \sum_{k' \neq k} \exp(\eta_{dtk'})$, $\kappa_{dtk} = N_{dtk} - \frac{N_{dt}}{2}$, and N_{dtk} is the number of elements to which topic k assigns in the contents posted by the user d at t. The first line represents the categorical likelihood, and there are no conjugate priors with the normal distribution. However, by introducing an auxiliary variable following the Pólya-Gamma distribution $\zeta_{dtk} \sim PG(N_{dt}, 0)$, we can transform the categorical likelihood into the normal kernel (Polson, Scott, and Windle, 2013) in the last line. Therefore, the conditional posterior distribution of η_{d1k} can be derived as follows.

$$\begin{aligned} &p(\eta_{d1k} \mid z_{d1}, \pi_{d1k}, \rho_{d1k}) \\ &\quad\propto\ p(z_{d1} = k \mid \eta_{d1k}) p(\eta_{d1k} \mid \pi_{d1k}, \rho_{d1k}) \\ &\quad\propto\ \exp\left\{-\frac{1}{2}\left(\zeta_{d1k} + \frac{1}{\rho_{d1k}^2}\right)\eta_{d1k}^2 + \eta_{d1k}\left(\kappa_{d1k} + \zeta_{d1k} \log \sum_{k' \neq k} \exp(\eta_{d1k'}) + \frac{\pi_{d1k}}{\rho_{d1k}^2}\right)\right\} \\ &\quad=\ N(\mu_{d1k}, \sigma_{d1k}^2). \end{aligned} \tag{3.7}$$

We calculate in the same way after $t = 2$ to obtain the filtering distribution as the normal distribution $p(\eta_{dtk} \mid \eta_{d1:t-1\backslash k}, \ldots) \propto N(\mu_{dtk}, \sigma_{dtk}^2)$, where $\sigma_{dtk}^2 = \left(\zeta_{dtk} + \frac{1}{\rho_{dtk}^2}\right)^{-1}$ and $\mu_{dtk} = \sigma_{dtk}^2 \left(\kappa_{dtk} + \zeta_{dtk} \log \sum_{k' \neq k} \exp\left(\eta_{dtk'} + \frac{\pi_{dtk}}{\rho_{dtk}^2}\right)\right)$.

Next, we derive the smoothing distribution as follows.

$$
\begin{aligned}
& p(\eta_{dt-1k} \mid \eta_{dtk}, \eta_{d1:t-1\backslash k}, \{\eta_{f1:tk}\}, \ldots) \\
& \quad \propto \ p(\eta_{dtk} \mid \eta_{dt-1k}, \{\eta_{ft-1k}, \ldots\}) p(\eta_{dt-1k} \mid \eta_{d1:t-1\backslash k}, \{\eta_{f1:t-1k}\}, \ldots) \\
& \quad \propto \ N(\tilde{\mu}_{dt-1k}, \tilde{\sigma}^2_{dt-1k}) \\
& \qquad \tilde{\sigma}^2_{dt-1k} = \left(\frac{\alpha_k^2}{\sigma_\epsilon^2} + \frac{1}{\sigma^2_{dt-1k}} \right)^{-1}, \\
& \qquad \tilde{\mu}_{dt-1k} = \tilde{\sigma}^2_{dt-1k} \left(\frac{{\alpha_{dk}}^2}{\sigma \ \epsilon} \left(\eta_{dtk} - \sum_{f \in F_d} \beta_{dfk} \bar{\eta}_{ft-1k} - \gamma_{tk} \right) + \frac{\mu_{dt-1k}}{\sigma^2_{dt-1k}} \right).
\end{aligned}
\tag{3.8}
$$

Therefore, in the estimation procedure, we calculate the mean μ_{dtk} and the variance σ^2_{dtk} using the filtering distribution in the forward ($t = 1, \ldots, T$), and then sample the topic distribution θ_{dtk} from the smoothing distribution in the backward ($t = T, \ldots, 1$). Also, from the findings of Polson, Scott, and Windle (2013), the posterior distribution of the auxiliary variable ζ_{dtk} is Pólya-Gamma class, $p(\zeta_{dtk} \mid \eta_{dtk}) = PG(N_{dt}, \psi_{dtk})$. The all posterior distributions of the parameters of the proposed model, including those of the other parameters derived in the Appendix A.2, are the form that can be sampled by the Gibbs sampler and the Metropolis-Hastings sampler.

3.5 Empirical Analysis

3.5.1 Dataset

In this section, we show the results of empirical analysis using the Pinterest [1] data to apply the proposed model to read social media data. Pinterest is a social networking site where users save images discovered on the Internet, as website bookmarking. The users also check the activity of other users they follow to see what images they have saved recently, that is, their ideas and interests. Therefore, social influences are expected to occur, such that users change their behaviors by being inspired by influential friends' unique ideas and interests. Furthermore, the influence may not be constant concerning the contents of the images. Users save images with several topics, such as fashion, interior design, and lifehack, on the Pinterest, and the social

[1] `https://www.pinterest.com/`

influence of each image topic may vary, for example, if a user's image collection about fashion topic is sophisticated, and the user may be influential for fashion topic, while if the same user's other collection, e.g., lifehack, is miscellaneous, and the user may have little influence on the followers for lifehack topic.

In collecting the Pinterest data, we first sampled users in a snowball-sampling fashion starting with one user, and then filtered them by conditions such as the activity during the observation period, and as a result, we selected 3,356 users. Observing the following relationship between them, (a total of 1,787 relationships), we found that the density of this network is 0.02% $(1,728/(3,356 \times 3,356))$.

We also collected the images saved by above users for two years (104 weeks) from December 2017, and the total of images are 206,215. To enable the proposed topic model to handle them, we used the Google Vision API [2] to detect objects in the image data. As a result, we detected 9,019 objects, and all image contents in the dataset were translated to multisets of the name of the objects.

3.5.2 Model Comparison

In this section, we demonstrate the statistical effectiveness of considering social influence between users in modeling social media content. To evaluate the features of the proposed model (Model 4), we consider three comparative models: the conventional LDA (Model 1), which does not consider time dependency in the process of image content generation, that is, $z_{dtn} \sim categorical(\theta_{dt})$ and $\theta_{dt} \sim Dirichlet(\theta_0)$, the dynamic linear topic model (DLTM) with random walk (Model 2), which considers the evolving topic distribution in a random walk fashion, that is, $z_{dtn} \sim categorical(\theta_{dt})$, $\theta_{dt} = softmax(\eta_{dt})$, and $\eta_{dtk} \sim N(\eta_{dt-1k}, 1)$, and the DLTM without social influence (Model 3), which considers the evolving topic distribution by self-influence and time-specific effect but does not consider social influence, that is, $\eta_{dtk} \sim N(\alpha_{dk} \cdot \eta_{dt-1k} + \gamma_{tk}, 1)$. These four models are used to estimate the 90% of dataset for each user and to compare the model performance on the remaining 10% out-sample data from two statistical perspectives, WAIC and Perplexity.

Table 3.1 reports the calculated measures for each model and number of topics (bold numbers in the table represents the best model for each number of topics).

[2] `https://cloud.google.com/vision/`

These result show that the proposed model outperforms the others in the both comparisons. This indicates that existing social influences between users in the dataset and modeling time dependency of topic distributions while considering social influence are statistically supported.

In the next section, we will discuss the estimated results of the proposed model (Model 4) with three topics because it performs the best in the both comparisons and over three comparable models.

TABLE 3.1: Values of WAIC and Perplexity for each model

	Model	Number of Topics							
		3	4	5	6	7	8	9	10
WAIC	1 (LDA)	22269.61	31407.33	41067.11	51203.32	59775.44	71193.31	82823.51	92490.85
	2 (DLTM w/ RW)	**21759.08**	31527.49	**38376.20**	51004.31	**58012.97**	71994.71	79954.15	93247.47
	3 (DLTM w/o SI)	22165.77	31348.61	39884.46	50045.11	59790.12	69347.65	80241.29	**88010.03**
	4 (DLTM w/ SI)	21979.58	**30941.84**	41498.24	**49198.22**	59610.81	**67436.17**	**77859.92**	89459.70
Perplexity	1 (LDA)	329.77	460.99	593.24	726.36	858.64	990.23	1125.02	1259.65
	2 (DLTM w/ R.W.)	322.85	470.48	630.76	758.97	928.19	1109.82	1267.18	1449.82
	3 (DLTM w/o S.I.)	300.84	405.33	510.39	617.42	757.12	**835.24**	945.51	**1057.59**
	4 (DLTM w/ S.I.)	**299.51**	**405.04**	**510.11**	**616.17**	**725.08**	835.60	**944.80**	1058.22

3.5.3 Estimation Results

In this section, we discuss about the estimated results. First, we show the estimated topics to understand what kind of image the users have saved in the Pinterest platform. Table 3.2 shows the top 10 object names for each topic. These objects have highest values in the estimated element distribution ϕ_k for each topic; therefore, we can interpret the meanings of topics from sets of these objects. Topic 1 represents the food topic related to images by which cuisine recipes storage techniques in the kitchen are showed, for example,*food*, *cuisine*, and *dish* are the related objects to Topic 1. Topic 2 represents the art & fashion topic about text art and clothing fashion, for example, *font*, *text*, and *clothing*. Topic 3 represents the interior design topic about furniture placement and wall color design, for example, *Room*, *Furniture*, and *interior design*.

Next, we see the dynamics of the topic distributions for each user, and Figure 3.1 shows the estimated topic distributions for first nine users. As shown in Figure 3.1, the topic distributions tell us what topics the user is interested in and how those interests are shifting over time. For example, user 1 and 4 have interests on two topics, and these interests change over time while conflicting each other. Also, users 5 and 8 were interested in the art & fashion topic at the beginning of the observation period, but over time, they became more interested in the interior design topic and eventually in the completely different food topic. On the other hand, user 2 has always been interested in the art & fashion topic only. In the following, we will see the hierarchical structure of the topic distributions that indicate whether their dynamics are influenced by self-influence, social influence, user-specific effects, or time trend.

Because of the model identification condition, the number of dimensions for each parameter is one minus from the number of topics. First, the estimates of self-influence parameter are shown in the above of Figure 3.2. These values indicate the users' original directions of interests, and the higher these values are, the more these users save the contents of a certain topic regardless of what their friends save or overall time trend. However, neither topic 1 nor 2 differed much in the shape of the histogram and the estimated values were on both sides of zero; therefore, no overall differences between three topics including topic 3 were observed from this parameter.

Then, the below of Figure 3.2 shows the time-series plots of the estimated time-specific random effects (γ_t), and these values indicate the overall time trend for each topic. These results indicate that the art & fashion topic has a more stable and popular trend compared with the food topic, whereas the trend in the food topic has been drastically changing, with a sharp rise at the end of the observation period. From the results, we can see that drastic changes may occur in the food topic among users as seen in the dynamics of the topic distributions for users 5 and 8 in Figure 3.1.

Finally, we discuss the results of the estimated social influence, which is the main objective of this study. Figure 3.3 shows the histogram of the estimated social influence. These results indicate that the estimated social influences are concentrated around zeros, where the shrinkage prior introduced in Section 3.4.2 is reflected. However, because the social influences are estimated at different values for each pair of users and each topic, it is difficult to summarize the whole results into a single figure. Here, we present sub-graphs of the Pinterest network weighted by the estimated social influence for each topic in Figure 3.4. This figure shows that the same pair of users may have different extent of social influence for different topics. For example, the influences between user 1326 and the surrounding users (1914, 938, and 2359) in the upper parts of Figure 3.4 represents the effect of the image contents saved by the surrounding users on the contents generating behaviors of the user 1326. The estimated influences of the art & fashion topic are larger than those of the food topic. Therefore, the surrounding users can be influentials on the art & fashion topic for user 1326. Also, user 1326 have many friends (users followed by the user 632), and the estimated influences of the friends on the user 1326 are greater for the art & fashion topic than for the food topic as a whole. It represents that the user 1326 tends to be more sensitive to the art & fashion topic.

However, the absolute magnitudes of social influence are very small, which indicating all influences are too concentrating around zero. We may need to modify the hyperparamters for the shrinkage priors introduced in Section 3.4.2. Alternatively, testing another formulations for shrinkage prior may also be valuable, for example, Bayesian lasso (Park and Casella, 2008) and horseshoe prior (Carvalho, Polson, and

TABLE 3.2: Top 10 object names with the highest value of element distributions of the proposed dynamic topic model

Topic 1: Food	Topic 2: Art & Fashion	Topic 3: Interior Design
Food	Font	Room
Cuisine	Text	Furniture
Dish	Clothing	Interior design
Ingredient	Art	Property
Produce	Fashion	Table
Recipe	Pattern	Floor
Dessert	Dress	Building
Baked goods	Illustration	Wall
Comfort food	Fashion accessory	Plant
Vegetable	Photography	Home

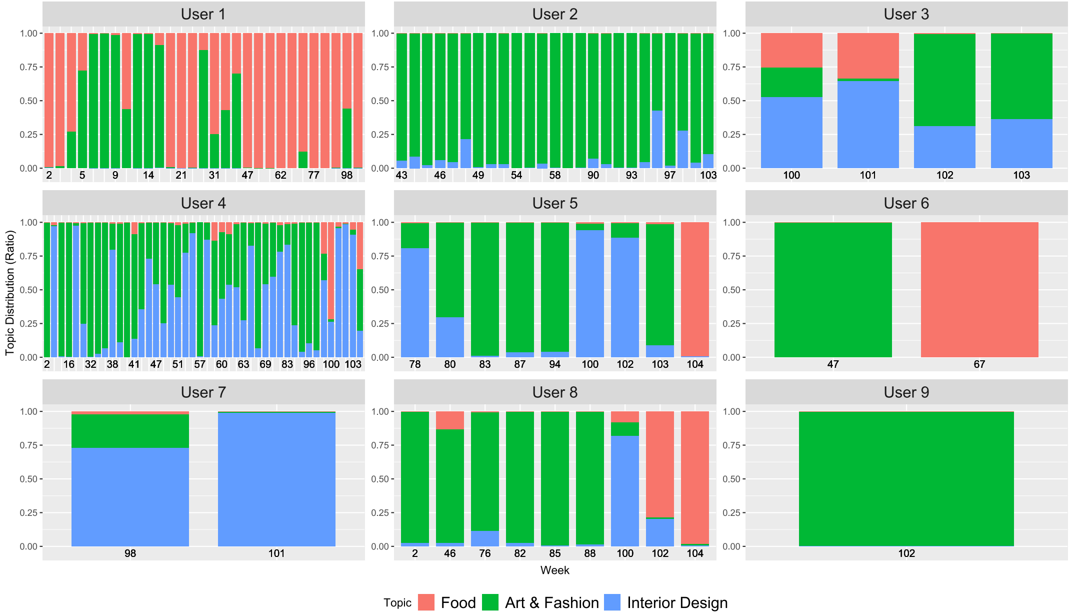


FIGURE 3.1: The estimated topic distributions for 9 users

Scott, 2009). These candidate prior distributions will be tested in future work.

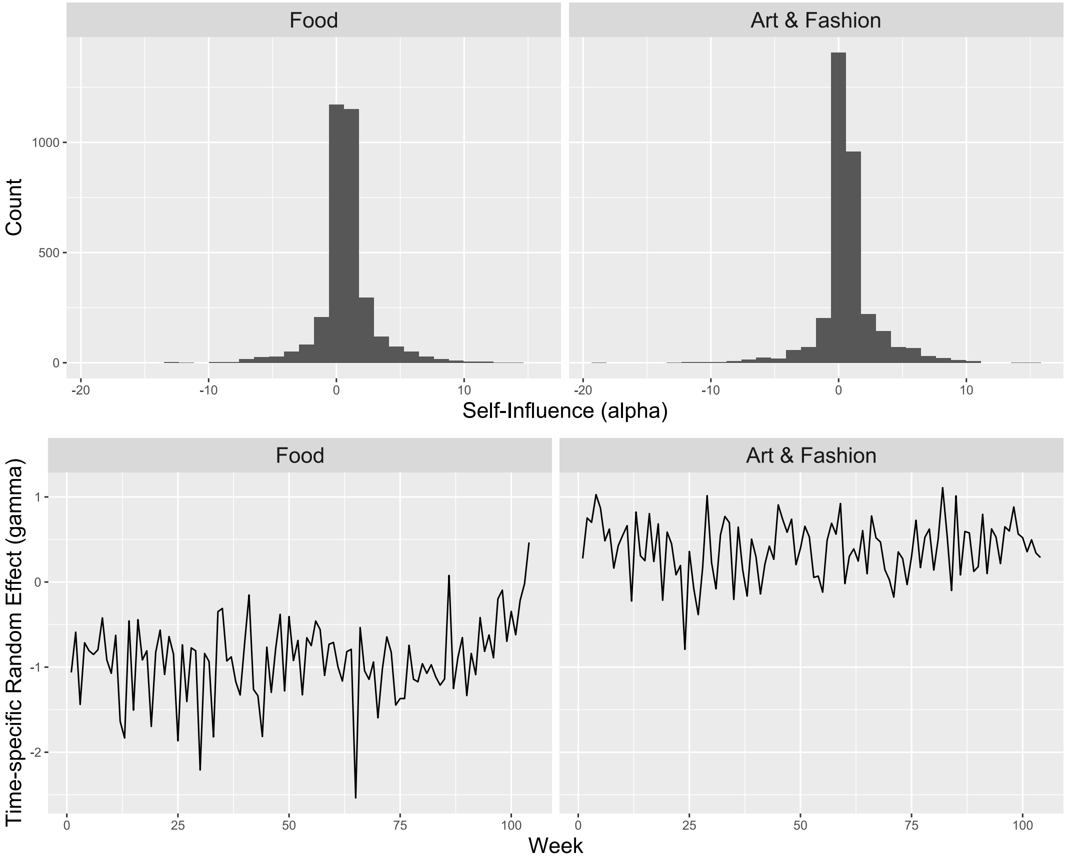


FIGURE 3.2: The histogram of the estimated self-influence (above) and the time-series plots of the estimated time-specific random effects (below)

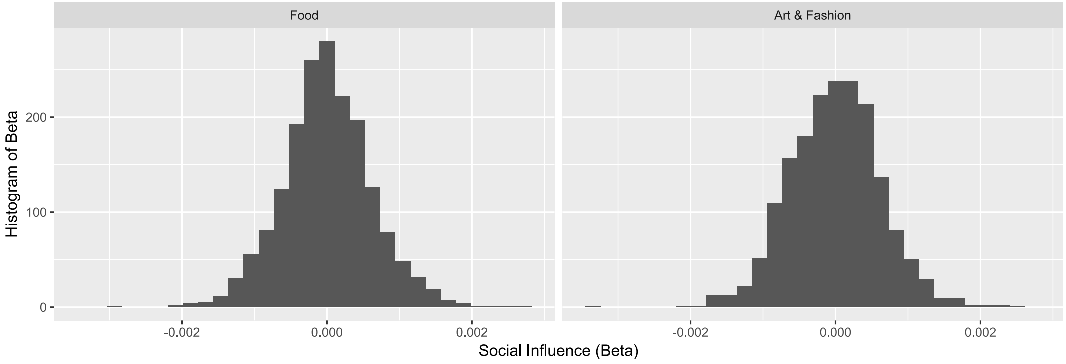


FIGURE 3.3: The histogram of the estimated social influence

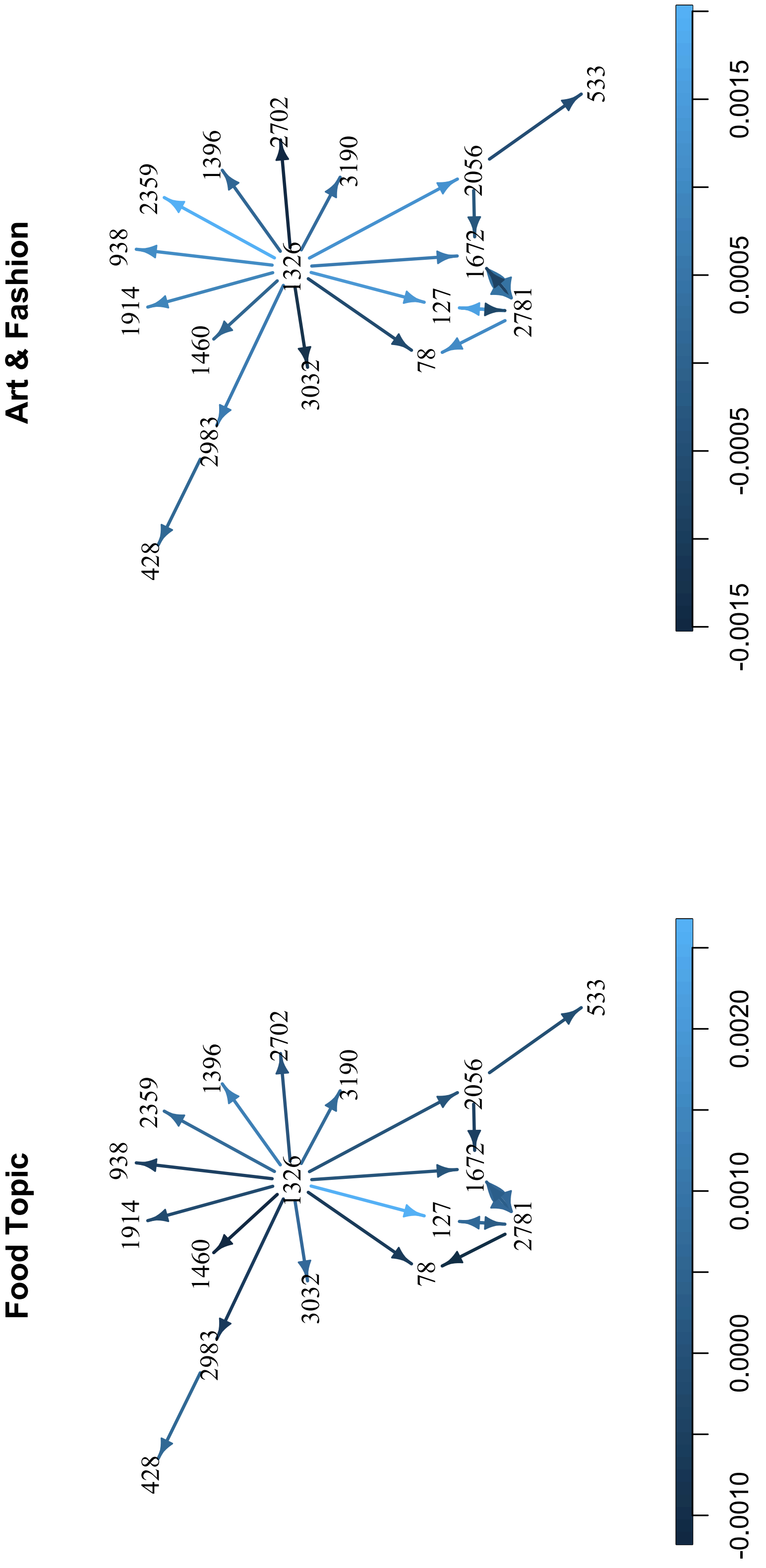


FIGURE 3.4: The part of network weighted by the estimated social influence

3.6 Conclusion

This study introduced a dynamic topic model for estimating social influence for each topic of user generated contents on social media. Most of the previous studies on social influence have focused only on the quantitative heterogeneity of the influence and neglect that the influence can vary for contents of the people behaviors even between the same pair of the people. This study contributes to the literature by proposing a model that captures how the topics of contents generated by users on social media are influenced by the content generation behaviors of the friends whom they follow, as the topic proportions themselves change over time.

The outstanding features of this study are that first, we estimate the social influence of the behaviors of generating unstructured data, such as text and images, which are often found in user generated contents on social media. In contrast, previous studies have estimated the influence of the purchasing behaviors of products. Moreover, the social influence taking into account the contents of the behaviors must be estimated using different approaches to the modeling on the structural data. Another feature is that the proposed model simultaneously estimates the topic of user generated contents and social influences varying for the topics. The previous studies should determine the dimensions of the product characteristics (e.g., the hedonic or functional features of the product) to estimate the influence of each dimension with those dimensions fixed. In our social media environment, however, the dimensions of the contents with various topics are difficult to determine using only prior knowledge. This study applies a topic modeling approach to estimate topics in the contents and clarify the social influences on their topic structure.

The empirical analysis demonstrates the usefulness of the proposed model for estimating topic-specific social influences on real social media. The proposed model reveals some explicit cluster structure in image data with various topics, such as food, art & fashion, and interior design, generated on the media, and estimates the dynamic changes in the topics of the generated contents. In addition, the proposed model estimates the social influences on the dynamic change in the topic proportions, distinguishing it from the sustained effects of the users' own topics and the user- and time-specific effects. As a result, we reveal topic-specific influences such as different influences on each topic even for the same user pair and users who are sensitive for a certain topic.

Further work may include the validation of the estimated topics and topic-specific influences. The proposed model clarifies the latent topic structure of UGC, and the estimated social influences only represents the relationship of content generation behaviors among users in latent dimensions. Therefore, without justification for the revealed latent topic structure, we cannot show the correctness for the social influences which are estimated on the latent structure. Previous studies have investigated how well their models correctly identify influencers against the comparative models by conducting additional simulation experiments. The consideration of such additional experiments to validate the results of this study is left for future work.

Chapter 4

The Effect of Manageable Perceived Topics in Customer Reviews on Product Satisfaction and Review Helpfulness

4.1 Introduction

Given the increase in the scale and scope of electronic commerce, online retailers such as Amazon, Walmart, and Taobao have experienced growth in the number of users making purchases on their online platforms. Most online retailers feature customer feedback systems in the form of customer reviews, including satisfaction scores (also called product ratings), textual reviews, and perceived helpfulness. These are useful to firms and marketers for understanding whether or not consumers prefer certain products, how consumers feel about a brand, the attributes relevant to decision-making, and other brands that fall into the same consideration set (Berger et al., 2020).

Identifying the perceived product attributes from user review content and recognizing their importance for customer evaluations is useful. An empirical study by Ghose, Ipeirotis, and Sundararajan (2007) argued that customer evaluations provide a meaningful basis for determining important product attributes that are central to marketing problems. Traditionally, the identification of such product attributes has been conducted with collected data from customer surveys and questionnaires (Fischer et al., 1999; Hoeffler, 2003). This requires the specification of a predefined and firm-oriented set of attributes that is selected by product designers and manufacturers, which is usually based on a limited amount of data due to the high cost of conducting labor- and time-intensive surveys.

Customer reviews consist of "the voice of the customer" and we can easily collect them without incurring any costs. Over the last decade, researchers have explored various methods for extracting product attributes from customer reviews and applying them to marketing research, e.g., market analysis (Lee and Bradlow, 2011; Tirunillai and Tellis, 2014). The customer reviews not only describe customer evaluations and their experiences with a product, but also provide insight regarding potential customers who read reviews to make future purchasing decisions. Chen and Xie (2008) suggested that online reviews help novice consumers identify products that best match their specific preferences. They concluded that, without reading reviews, novice consumers might be less likely to buy a product if the seller-created product attributes were only available to them. Obviously, consumers prefer to read customer reviews before making purchase decisions to reduce their perceived risk in buying a product and recognize such user-generated content as rather trustworthy by sharing their views. In the online review system considered in this study, review readers evaluate reviews, and then, they vote when they feel that it is helpful. Our model considers the interaction between review writers and readers to explore the effect of satisfaction rating scores on the number of votes for helpfulness.

To analyze customer reviews, we first extracted the perceived product attributes mentioned in the reviews. In the existing literature, several frameworks for understanding product attributes in online customer reviews have been proposed (e.g., Decker and Trusov, 2010). Most of these studies adopt a rule-based approach that translates words or phrases into product attributes on a one-to-one basis. They create lexicons for this translation using humans or useful tools such as machine learning and then map words or phrases to product attributes using the lexicons. Then, they construct a model to explain the relationship between quantified concepts from the reviews and dependent variables such as the rating score, which represents customer satisfaction.

The usefulness and advantages of this one-to-one translation of words to attributes is limited in three ways. First, it is difficult to simultaneously achieve a high level of precision and a low cost. If we use a generic lexicon to keep production costs low, we cannot accommodate specific words and phrases in the domain, making it more difficult to correctly convert them into product attributes. By contrast, creating a lexicon for each domain would be too labor- and time-intensive. Second, a tremendous amount of review data is produced daily and therefore it is not possible to create a lexicon accounting for all word trends in the reviews. Third, one-to-one translations cannot deal with polysemous words such as *tie* and *book*. These words represent different meanings according to their specific context in each review. To deal with polysemous words, we need to carefully examine their co-occurrences with surrounding words.

Topic models are able to address these limitations. They assume that each word might be assigned to multiple topics (i.e., perceived attributes) according to its context and thus these perceived attributes can be flexibly extracted from the review text. They were originally used to extract latent semantic meaning from a large text corpus and classify documents and predict new documents. Many researchers have proposed various efficient estimation methods for the big data online environment in which text data are accumulated and updated (e.g., Hoffman, Bach, and Blei, 2010). In addition, this approach involves little human intervention. Since we do not need to know the latent product attribute dimensions in advance, human error and bias is minimized.

We employ a representative topic model known as the latent Dirichlet allocation (LDA) model put forth by Blei, Ng, and Jordan (2003), in which no word is given to any topic; words are assigned to the most likely topic through a learning process. This produces a set of words characterized by cohesion that is incomprehensible to humans. As a result, the extracted topics, i.e., the perceived attributes, are often not interpretable, as discussed by Mimno et al. (2011).

To address this problem, we propose a partially labeled topic model that provides symbolic words representing product attributes with some topics in advance and leaves the remaining topics unspecified. Regarding product price attributes, for instance, the words "expensive" and "cheap" can be viewed as representative words. The topic assignment of these words is fixed in advance and other topics are kept free when applying the topic model. At the same time, these topics are extracted so as to explain the satisfaction score of review writers and the helpfulness of readers, respectively, by using supervised modeling. The labeled and supervised topic-based response functions of the satisfaction score and helpfulness count are connected to obtain an integrated model that can accommodate the interaction discussed above. We naturally incorporate prior knowledge into the model for the sake of topical interpretability, at the cost of model fit. We examine its costs and benefits in an empirical analysis to demonstrate the usefulness of our proposed model.

In this empirical study, we use Amazon customer review data on a potato chip product and compare the performance of our model with other existing models on the points of the interpretability of the extracted topics. We demonstrate that model fit and the predictive performance of the word labeling restricted model is comparable with that of nonrestricted models, and that our model has the added advantage of providing interpretable and manageable perceived attributes for use by marketers. The parameter estimates provide useful findings such as the fact that the "ingredient" topic in reviews decreases the level of the satisfaction score and perceived helpfulness to readers. Conversely, the "health" topic increases the levels of both.

The rest of this chapter is organized as follows: In Section 4.2, we discuss related studies in the relevant body of the existing literature. In Section 4.3, we describe the details of the dataset used in this study and how to construct "labeled" topics. Then, in Section 4.4, we propose a partially labeled supervised LDA model. Section 4.5 presents the model's empirical application to Amazon customer review data and presents a discussion. Finally, we provide concluding remarks and directions for future research in Section 4.6.

4.2 Literature Review

4.2.1 Customer Review Analysis for Understanding Their Impact on Satisfaction and Helpfulness

Researchers in marketing and management propose several approaches to extracting the product attributes of customer reviews and obtaining managerial insight through their application. Many studies used customer reviews in the existing literature. For instance, Lee and Bradlow (2011) and Netzer et al. (2012) explored market structure by estimating the relationship between product attributes and customer satisfaction or sales. Ghose, Ipeirotis, and Li (2012) used the extracted perceived attributes to improve search algorithms for product web pages.

Decker and Trusov (2010) and Xiao, Wei, and Dong (2016) discussed the impact of perceived product attributes that were positively and negatively mentioned in product ratings using the latent class Poisson regression model and a heterogeneous multinomial choice model. Archak, Ghose, and Ipeirotis (2011) proposed a demand model that captured the effect of product attributes on sales by considering the heterogeneity of each word that qualifies the attributes. While these studies treated all reviews as subjects for analysis, Qi et al., 2016 used human power and machine learning methods to select only useful reviews and analyzed the less biased effects of attributes on product ratings.

The customer reviews not only contain assessments of product attributes by customers who already purchased the product but also affect the perceived helpfulness of consumers as potential buyers. Some studies have analyzed the types of reviews that are helpful to review readers. For example, Chen and Xie (2008) explored the interactive effect of seller-created product attributes and buyer-created review information on the usage experiences of readers. They claimed that customer reviews helped consumers identify the specific products that best matched their idiosyncratic usage conditions.

Yin, Bond, and Zhang (2017) and Felbermayr and Nanopoulos (2016) examined the effect of emotional arousal in the reviews on readers' helpfulness by accounting for the variations in the exact degree of emotional arousal. Mudambi and Schuff

(2010) revealed the factors that make reviews helpful to customers during the purchase decision-making process and described the effect of review ratings, review depth, and product search.

4.2.2 Extracting Product Attributes from Customer Reviews

The various approaches to analyzing customer reviews reveal the relationships between the product attributes mentioned therein and specific dependent variables such as satisfaction, sales, and reader helpfulness. They follow a two-step process that extracts product attributes from review text and then constructs a model to estimate the effect of the product attributes on the specified dependent variables. The literature on the first step, extracting the product attributes from the review text, is divided into two approaches: the rule-based and the model-based approach.

Most studies that take a rule-based approach map words and product attributes on a one-to-one basis. After preprocessing review text, including removing stop words and word stemming, they create rules that determine perceived product attributes based on words identified using human power (Moon and Kamakura, 2017; Hou et al., 2019) or machine learning techniques such as clustering (Lee and Bradlow, 2011; Archak, Ghose, and Ipeirotis, 2011). To identify the correspondence between these words and perceived attributes, unstructured textual data are transformed into quantified variables and the effects of the attributes on dependent variables are explored through either a regression-based model or a choice model. However, these rule-based approaches ignore the fluctuations in meaning and sentiment of words according to their specific context.

To capture these fluctuations, the model-based approach that uses topic modeling, specifically LDA (Blei, Ng, and Jordan, 2003), has been applied in the literature. The LDA model reflects the generative process of text in which a word is assumed to be generated from the vocabulary when its latent topic is given. In addition, topics might differ according to context, even about the same word. In customer review analysis, Tirunillai and Tellis (2014) proposed an extended LDA model by adding a mechanism that incorporated the latent sentiment of words. They extracted the key latent dimensions of consumer satisfaction to conduct brand-positioning analysis. Büschken and Allenby (2016) extended a topic model to consider sentence-based topics rather than just independently assigning words to topics and discussed the importance of each product attribute to customer satisfaction.

However, extracted topics are not always well-interpreted using conventional LDA models because some words with less semantic cohesion might be clustered in a certain topic. In the fields of machine learning and natural language processing, several studies have developed metrics for evaluating the quality of extracted topics (often called *cohesion* scores), which measure the degree of topic interpretability for humans (e.g., Chang et al., 2009; Mimno et al., 2011). They argue that topics extracted using popular topic models, such as probabilistic latent semantic indexing (pLSI, Hoffman, 1999) and correlated topic models (CTM, Blei and Lafferty, 2005) and the LDA model, are not always semantically meaningful. In this sense, these models are generally limited in the manageability of their extracted attributes because we cannot obtain useful insight into the structure of customer satisfaction and helpfulness when such uninterpretable topics are plugged into the response function as covariates.

In this study, we first employ a partially labeled topic model that provides several representative words with product attributes in advance and assigns topics to each word based on these seed words. Similar approaches using seed words or seed labels have been previously proposed in the literature. Lin et al. (2012) and Tirunillai and Tellis (2014) used seed words to estimate word sentiment by fixing some word sentiments to a polarity, such as *good* and *great* as positive words, or *bad* and *terrible* as negative words. Ramage et al. (2009) employed some tags previously assigned to documents as seed topics; then the word topics in the documents were determined from the set of seed topics.

Some studies have integrated the textual information quantified by the LDA model into response function models, such as linear regression and multinomial probit models, to treat the topic extraction and the response function as two distinct models (Qi et al., 2016; Bi et al., 2019), whereas others treat them together as a single model (Büschken and Allenby, 2016; Puranam, Narayan, and Kadiyali, 2017) by using the supervised LDA model (Blei and McAuliffe, 2007).

We combine the partially labeled LDA and supervised LDA models to jointly estimate the features of perceived product attributes and their relationships with product satisfaction and reader helpfulness in customer reviews using a one-step procedure.

4.3 Data

In this section, we explain the details of the dataset used in this study and construct "labeled" topics. We use Amazon customer review data collected by the authors consisting of 1,178 customer reviews of a specific potato chip product that was sold on the Amazon platform between March 2009 and October 2019. Selecting a single product facilitates the assumption of perceived topics on the point of product features.

This dataset includes variables including review text, product rating score, voted helpfulness count, and control variables. Specifically, after buying and accepting goods, customers are allowed to post their feelings and experiences by writing and posting a textual review to the Amazon website product page. At the same time, these customers can also grade products by assigning a "star" rating between 1 (lowest) and 5 (highest). A reader of this online review can then evaluate whether the review was helpful in his or her product consideration and purchasing decision by clicking on the "helpfulness" button and rating it as either helpful or unhelpful. As control variables, we consider four types of status badges: purchase verification, top contributor, top reviewer, and vine voice. Purchase verification indicates that the reviewing customer was verified as having purchased the product being reviewed on the Amazon platform. The top contributor badge is awarded to customers who frequently share reviews and answer customer questions. Top reviewers are identified by Amazon's reviewer rankings, although the ranking system algorithm has not been disclosed. Vine voice is an invitation functionality that gives Amazon reviewers advanced access to not-yet-released products for the purpose of writing a review. Therefore, these variables might also impact product ratings and reader attitudes, as well as perceived topics. In the next section, we construct a model to examine and discover these relationships.

Before data analysis, the review text was preprocessed. First, for each document, we split the text into word sets and substituted capitalized letters with lower case

TABLE 4.1: List of labeled words for each product attribute

Flavor and Taste:	salt, vinegar, flavor, cheddar, taste, bbq crunchy, texture, tasty, sweet, pepper, salty
Packing:	case, box, product, store, bag, pack, weight
Healthy:	healthy, fat, calories
Money and Buying:	try, buy, get, bought, find, price
Ingredients:	oil, flour, ingredients, starch, flakes

letters. Next, we excluded numbers, punctuation, and popular stop words (e.g., a, the, I, they). We also transformed inflected words to their word stems to reduce redundancy in the created corpus. Finally, we excluded frequently used and rare words because they adversely affect the extraction or interpretation of topics after extraction. As a result, the created corpus included 716 unique words and each review consisted of an average of 9.8 words.

The purpose of this study is to extract manageable topics from textual reviews and explore the interconnections between topics, review writer satisfaction, and review reader perceived helpfulness. This is especially important for online retailers and marketers in developing and improving their products according to the specific relationships discovered between objective product features and user satisfaction or perceived helpfulness.

To extract manageable topics, we assumed a certain number of perceived topics on the features of food. The literature mentions several features of food such as culinary quality (Chi et al., 2013), taste and flavor (Andersen and Hyldig, 2015), service and price (Mason and Nassivera, 2013), health, nutrition, and low-fat ingredients (Küster and Vila, 2017). At the same time, to identify the possible features of food in the dataset, we evaluated subsets of customer reviews and observed frequently used words in the entire vocabulary. Based on these observations, the features of food were related to flavor, taste, packing, weight, food ingredients, cooking methods, and purchases. According to previous related literature and the observed product features in the dataset, we combined synonym topics and proposed five perceived topics: (1) flavor and taste, (2) packing, (3) healthy, (4) money and buying, and (5) ingredients. In this study, we also selected seed words for each labeled topic; these words should be symbolic and representative of each product feature. Table 4.1 provides a list of labeled words.

4.4 Model

4.4.1 Partially Labeled and Supervised Topic Model

Below, we define the partially labeled supervised LDA (PLS-LDA) model, which combines the labeled topic and supervised topic models.

First, we show how to determine the labeled topics using labeled words. After removing unnecessary words, we consider the total vocabulary, which included all words in customer reviews as text content. To make the topics interpretable, some representative words are selected from the total vocabulary of each topic, as discussed in the previous section; we call these words *labeled words* as follows. Simultaneously, the remaining words with no predefined labels are called *nonlabeled words* and assigned to topics according to a specific distribution.

To develop the specific vocabulary of the topic k from the total vocabulary available, we employ a methodology from the existing literature such as labeled LDA (Ramage et al., 2009) or joint-sentiment LDA (Lin et al., 2012; Tirunillai and Tellis, 2014). We first generate the topic's *transformation matrix* $\Lambda^{(k)}$ $(V_k \times V)$, conditional on $\lambda_1, \lambda_2, \ldots, \lambda_{V_k}$, where $\lambda_i, i = 1, \ldots, V_k$, is the sequence of the number of labeled words for topic k and nonlabeled words, V_k and V represent the number of labeled words and nonlabeled words for topic k and the number of total vocabulary in the dataset, respectively. For each row $i \in \{1, 2, \ldots, V_k\}$ and column $j \in \{1, 2, \ldots, V\}$, we set

$$\Lambda_{ij}^{(k)} = \begin{cases} 1 & \text{if } \lambda_i = j \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

In this model, the topic-vocabulary vector ϕ_k, called the word distribution, is assumed to follow the Dirichlet distribution, $\phi_k \sim Dirichlet(\beta_k^*)$, where β_k^* is transformed from the V-dimensional hyper parameter β via the transformation matrix, $\beta_k^* = \Lambda^{(k)}\beta$. Hence, the number of dimensions of β_k^* is V_k.

The remaining LDA part of the model is not restricted and followed conventional LDA model procedure. That is, the topic assignment for the n-th word in review d which consists of N_d words without considering their order (bag-of-words assumption) is assumed to follow a categorical distribution, $z_{dn} \sim categorical(\theta_d)$, where θ_d

is a topic distribution that represents a topic proportion within the text of the review d and follows the Dirichlet prior distribution, $\theta_d \sim Dirichlet(\alpha)$. We assume that a word assigned to topic k is generated from the corresponding word distribution, $w_{dn} \mid z_{dn} = k \sim categorical(\phi_k)$.

Next, we develop the response function part of our model. We use two dependent variables: satisfaction score and helpfulness count. The satisfaction score reflects the current evaluation of customers based on their past experiences that already purchased the product and we assumed that the helpfulness count implied the interest in and expectations of the product by other consumers that might purchase it in the future. In addition, as discussed in the previous section, the labeled topics in customer reviews are directly built as covariates for the variations in satisfaction scores and as helpful product expectation references. For example, if one product feature is particularly satisfying to customer needs, we expect that the topic related to this feature co-occurs with a high satisfaction score in online customer reviews. Conversely, topics related to dissatisfying features should co-occur with low satisfaction scores. Regarding its connection with review helpfulness, a textual review is likely to be regarded as helpful by readers if it contains the topics in which they are interested.

Given the word-topic assignments, z_{dn}, $N_{dl} = \sum_{n=1}^{N_d} = \mathbb{I}\{z_{dn} = l\}$ is the number of words assigned to topic l, which we use as covariates after logarithmic transformation. Note that the number of covariates for these topic assignments is the same as the number of labeled topics, L (which is set to five in this study). However, the number of total topics can differ from the number of labeled topics. Let the number of total topics be K and the rest of the topics, $K - L$, are nonlabeled and do not include any labeled words. Nonlabeled topics are extracted from the review text but are not considered as covariates for both object variables for the sake of manageability.

The number of total topics, K, is determined by model selection criteria of the deviance information criterion (DIC, Spiegelhalter et al., 2002) and the widely applicable information criterion (WAIC, Watanabe, 2010), as demonstrated in Section 4.5. In addition to the labeled topic variables, some control variables also work as covariates. We use the following four status variables: purchase verification, top contributor, top reviewer, and vine voice. The status badges are displayed next to the user icon if the user qualified for that status. Word count variables, which represent the number of words included in the customer review, are also considered.

We assume that the satisfaction score, which was measured using a five-point scale, follows the ordered probit model and the helpfulness count with positive integers follow the Poisson regression model. First, let the satisfaction score of a review d be $y_{s,d}$, which follows the ordered probit model:

$$\begin{aligned} y_{s,d} &= r, \qquad \text{if } \tau_{r-1} \leq y^*_{s,d} < \tau_r \\ y^*_{s,d} &= \sum_{l=1}^{L} \gamma_{s,l} \cdot \log(N_{dl} + 1) + \sum_{m=1}^{5} \delta_{s,m} \cdot x_{s,dm} + \epsilon_d; \qquad \epsilon_d \sim N(0,1). \end{aligned} \tag{4.2}$$

In the above, the thresholds $\{\tau_r\}$ work for realizing discrete satisfaction scores through the latent continuous variable, $y^*_{s,d}$, and the thresholds τ_0 and τ_R ($R = 5$ in the Amazon dataset) are set to $-\infty$ and ∞, respectively. $x_{s,d}$ is a vector of the control variables– purchase verification, top contributor, top reviewer, vine voice, and word counts. The error term ϵ_d is assumed to follow a standard normal distribution and the model does not include the intercept term for identifying $R - 1$ thresholds.

Next, when the satisfaction rating score $y_{s,d}$ is given, we define the response model of the helpfulness count $y_{h,d}$ for review d using the Poisson regression model:

$$\begin{aligned} y_{h,d} &\sim Poisson(y^*_{h,d}) \\ y^*_{h,d} &= \sum_{l=1}^{L} \gamma_{h,l} \cdot \log(N_{dl} + 1) + \sum_{m=1}^{5} \delta_{h,m} \cdot x_{h,dm} + \delta_{h,6} y_{s,d}, \end{aligned} \tag{4.3}$$

where $x_{h,d}$ is common with satisfaction probit model, Equation (4.2) and $y_{s,d}$ is included by the findings of literature (e.g., Ho-Dac, Carson, and Moore, 2013; Mauri and Minazzi, 2013; Ludwig et al., 2013), which demonstrate that positive and negative online customer reviews affect reader purchase intentions and expectations. We also explore the effect of the level of customer satisfaction on the perceived helpfulness of readers.

Therefore, the satisfaction and helpfulness models in Equations (4.2) and (4.3) are sequentially connected by way of observation $y_{s,d}$ to form the integrated PLS-LDA model and its full-joint likelihood is described as follows:

$$\begin{aligned}
&p(W, Y_s, Y_h, X_s, X_h, Z, \theta, \phi, \tau, \alpha, \beta, \Lambda, \gamma_s, \gamma_h, \delta_s, \delta_h) \\
&= \prod_{d=1}^{D} \{p(y_{s,d} \mid z_d, x_{s,d}, \gamma_s, \delta_s, \tau) p(y_{h,d} \mid z_d, x_{h,d}, y_{s,d}, \gamma_h, \delta_h) \\
&\quad \left. \prod_{n=1}^{N_d} p(w_{dn} \mid z_{dn}, \phi) p(z_{dn} \mid \theta_d) \right\} \left\{ \prod_{d=1}^{D} p(\theta_d \mid \alpha) \right\} \left\{ \prod_{k=1}^{K} p(\phi_k \mid \Lambda^{(k)}, \beta) \right\} \\
&\quad p(\gamma_s, \gamma_h, \delta_s, \delta_h),
\end{aligned} \tag{4.4}$$

where $p(\gamma_s, \gamma_h, \delta_s, \delta_h)$ is the prior distribution of the model coefficients and the setting of distribution and hyper parameters is provided in Appendix A.3. To estimate the model, we applied a hybrid Bayesian estimation using semi-collapsed Gibbs sampling and the random walk Metropolis-Hastings method. The details of the estimation procedure are provided in Appendix A.3.

4.5 Empirical Analysis

4.5.1 Comparison Results

In this section, we compare the model with comparable alternative models from two viewpoints. The first is model selection by statistical criterion and the second is the manageability of extracted topics. Through these two comparisons, we examine how well the model performs and how manageably useful it is for marketers against alternative models, even at the cost of model fit. The details of the estimation settings are provided in the appendix.

We consider two comparative models, the separate model and the supervised LDA model, denoted as separate and S-LDA, respectively. The separate model contains two separate processes, namely, extraction of product attributes from review text using the LDA model and response model explaining satisfaction and helpfulness using ordered probit and Poisson regression models. The S-LDA model has no constrain related to labeled words. The parameters are estimated using Gibbs and Metropolis-Hastings sampling with the same settings as the PLS-LDA model.

First, we discuss the results of the model comparison using DIC and WAIC. The two criteria compare models from different perspectives, that is, DIC considers the model's goodness of fit and complexity, and the WAIC assesses the model's generalization error.

Figure 4.1 illustrates the value of DIC and WAIC as summations of those for submodels for product satisfaction and helpfulness in the range of the number of topics between 5 and 15 for the three models. Line colors identify the various models, the separate model (red), the S-LDA model (green), and the PLS-LDA model (blue), and the dots represent the smallest values among the variations in the number of topics for each model. The comparison between the separate and supervised models, S-LDA and PLS-LDA, shows that the separate model performs worse than the others for all numbers of topics based on both criteria. Whereas the separate and supervised models have been used in previous studies (e.g., Moon and Kamakura, 2017; Büschken and Allenby, 2016, respectively), supervised models are supported in terms of predictive measures by the information criterion.

A comparison of the PLS-LDA model with the S-LDA model shows that the S-LDA model is better overall, even though our PLS-LDA model has smaller DIC and WAIC values for certain numbers of topics. One possible explanation for this fact is that our PLS-LDA model has restrictions on labeled words and that the parameters with respect to these words are fixed at certain values and might deviate from values that achieve a better fit. In contrast, the S-LDA model is allowed to be well-optimized without any restrictions. However, we should note that, more importantly, the difference between both models is not very large, especially when compared with the WAIC values, which are almost the same for both models. This indicates that the labeled word constraints are suitable for explaining data variation. Then, we recognize that our model is better when considering the intrinsic benefits of topic interpretability and manageability for marketers.

Next, we compare our PLS-LDA model with the S-LDA model from the perspective of topic interpretability. Table 4.2 provides the top 15 words distributed for each topic in descending order. Asterisks next to words indicate labeled words. For a simple comparison, the number of topics for both models are 5, which is the same as the number of expected perceived topics determined before the analysis. The upper part of Table 4.2 shows that, for our model, five interpretable topics are attributable to the labeled words. In addition, our model assigns seemingly correct topics to

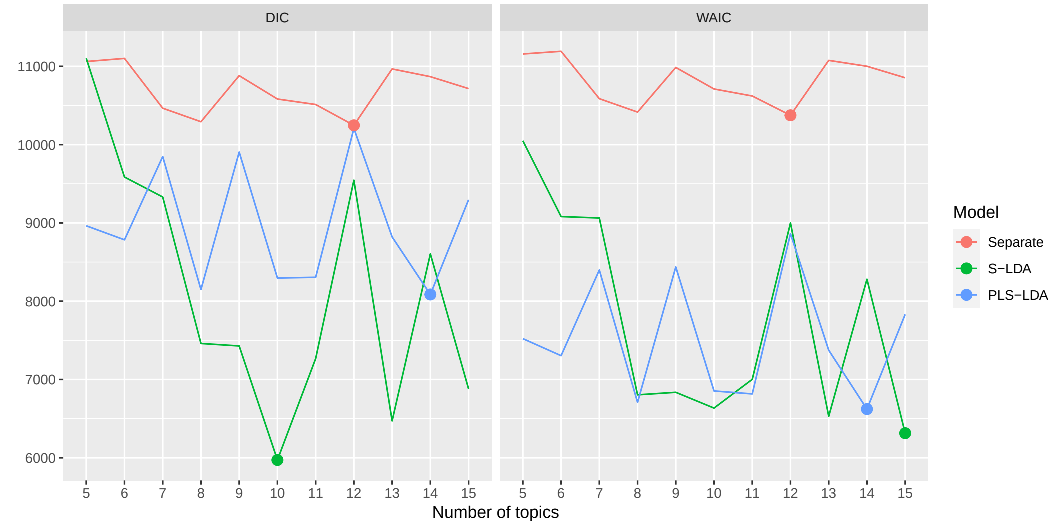


FIGURE 4.1: Values of DIC (left) and WAIC (right)

words that are related to the labeled words and therefore they are appropriate for making up the same topic. For example, *lime* and *chili* are in the flavor and taste topic, *size* is in the packing topic, and *diet* is in the healthy topic; these words are nonlabeled but appropriate for constituting the same topic along with the labeled words. In contrast, most topics in the lower part of Table 4.2 for S-LDA cannot be interpreted because of meaningless sets of words, for example, some inconsistent words, such as *weight*, *price*, and *calorie*, are in the same Topic 3, and the same words, such as *bag* and *flavor*, are related to multiple topics. Therefore, we face difficulties interpreting the meaning of topics using the S-LDA model.

In summary, these results demonstrate that our model is reliable in both explaining the variations in product satisfaction and reader helpfulness and the interpretability of topics by labeling certain seed words as representative of topics, even at the cost of model fit. In the next section, we discuss the estimation results of the model parameters.

4.5.2 Discussion of Estimation Results

Table 4.3 shows the results of parameter estimates when the number of total topics is 14, which is the smallest value of both criteria in Figure 4.1. Table 4.3 provides that the estimated posterior means of threshold parameters and coefficients of labeled topics and control variables. The asterisks next to the posterior mean indicate the

TABLE 4.2: The top 15 words of word distribution for each topic in descending order

partially labeled supervised LDA				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
(Flavor & Taste)	(Packing)	(Healthy)	(Money & Buying)	(Ingredient)
flavor*	bag*	calorie*	try*	ingredient*
taste*	box*	fat*	get*	oil*
salt*	case*	healthy*	buy*	rice
sweet*	product*	fry	find*	flour*
bbq*	pack*	delicious	bought*	definitely
vinegar*	store*	always	price*	corn
salty*	weight*	crave	order	kid
pepper*	size	per	origin	back
texture*	order	diet	know	thought
lime	whole	star	since	final
crunch	need	ever	bad	disappoint
crunchy*	watcher	recommend	healthier	kind
chili	variety	sodium	far	purchase
cheddar*	keep	lunch	greasy	say
tasty*	addict	look	addict	list

supervised LDA				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
sweet	bag	bag	flavor	flavor
fat	taste	find	taste	salt
taste	try	weight	try	vinegar
calorie	buy	buy	lime	taste
flavor	get	price	chili	try
rice	package	calorie	texture	bbq
healthy	box	store	expect	pepper
bag	order	healthy	though	origin
ingredient	pack	get	get	calorie
salty	bought	bbq	spicy	garlic
oil	purchase	size	product	cheddar
bad	never	order	need	order
flour	look	every	salt	prefer
fry	said	day	calorie	fat
sodium	people	keep	say	sour

significance of the parameter in the sense of 95% highest posterior density region (HPD).

First, the threshold parameters (τ_r) indicate that an approximately 0.50 increase in the latent continuous rating is associated with a one-point increase in the observed discrete rating. Because the covariates for the topics represent the log-transformed number of words assigned to the topic, the coefficients of the labeled topics (γ_l) are interpreted as a substantial change in the ratings for a 1% increase in the number of words for that topic. For example, if the number of words associated with the topic "flavor and taste" increases by 1%, the expected change in the latent rating is -0.58, translating to an almost one-point decline in customer satisfaction.

Table 4.3 also provides some interesting findings related to the model coefficients. The coefficients of Topics 1, 2, and 5 for the satisfaction model are estimated as negative and only those of Topic 3 are estimated as positive. This indicates that dissatisfied customers are more likely to talk about the "flavor and taste," "packing," and "ingredient" topics and satisfied customers are more likely to talk about the "health" topic. In addition, significant coefficients of the helpfulness regression are related to the health and ingredient topics. This result suggests that reviews including "health" topic words are regarded as helpful and reviews including "ingredients" topic words are less helpful.

Remarkable findings are also observed in the coefficients of control variables (δ_m). As the coefficients of purchase verifications are positively significant for both objective variables, customers who purchased the product can be viewed as being satisfied relative to customers who did not make a purchase, and readers find that reviews by customers who made purchases are more helpful. Similarly, more satisfied customers write longer reviews and such reviews are considered helpful by readers. The last finding is the negative effect of the satisfaction score on reader helpfulness. Readers tend to find critical reviews with low satisfaction scores more helpful than positive reviews by highly satisfied customers.

In conclusion, these results answer our research questions: "What is the relationship between product attributes and the satisfaction of customers who make purchases?" and "What attributes do readers of customer reviews expect to consider

TABLE 4.3: Estimation results of the proposed model

Parameters	Posterior Mean (95 % HPD interval)	
	Satisfaction Score	Helpfulness Count
τ_1	-2.095* (-2.340, -1.843)	—
τ_2	-1.635* (-1.808, -1.425)	—
τ_3	-1.102* (-1.272, -0.937)	—
τ_4	-0.587* (-0.761, -0.392)	—
γ_1 (Flavor & Taste topic)	-0.584* (-0.773, -0.404)	0.081 (-0.018, 0.182)
γ_2 (Packing topic)	-0.382* (-0.605, -0.160)	0.009 (-0.123, 0.139)
γ_3 (Health topic)	1.046* (0.797, 1.412)	0.238* (0.050, 0.439)
γ_4 (Money & Buying topic)	0.075 (-0.190, 0.338)	-0.114 (-0.274, 0.044)
γ_5 (Ingredient topic)	-1.612* (-2.147, -1.115)	-0.274* (-0.535, -0.019)
δ_1 (Purchase verification)	0.294* (0.073, 0.502)	0.234* (0.059, 0.410)
δ_2 (Top contributor)	0.678 (-0.852, 2.161)	1.004* (0.234, 1.773)
δ_3 (Top reviewer)	0.748 (-1.494, 3.128)	-0.489 (-1.682, 0.707)
δ_4 (Vine voice)	0.593 (-0.924, 2.043)	0.076 (-0.742, 0.931)
δ_5 (Word counts)	0.012* (0.006, 0.018)	0.004* (0.001, 0.007)
δ_6 (Satisfaction score)	—	-0.152* (-0.190, -0.113)

useful information?" These findings were obtained by labeling seed words and extracting interpretable perceived topics.

4.5.3 Marketing Values of This Study

Lastly, we clarify the marketing values of this study in this section. This study can be valuable for companies and managers from the following three contributions.

First, this study is based on the model-based customer review analysis, and once the model is constructed, it is possible to perform everything in an automation manner from the extraction of product attributes in the review text to estimation of the preference structure. Moreover, it is also possible to track dynamic changes through fixed-point observation. On the other hand, the rule-based approach, which makes a one-to-one correspondence between words and product attributes, enables highly accurate analysis through detailed attribute extraction. However, there are some inefficiencies, such as high cost of rule (dictionary) creation, low generalizability to other domain, and difficulty to take into account fluctuations where words have different meanings depending on the context. Incorporating model-based customer review analysis by the proposed model into their daily operations enables companies can gain a detailed understanding of consumer feedback without any additive costs.

Second, the proposed partially-labeled and supervised LDA model can improve the topic interpretability while the existing models applying conventional LDA model to extract product attributes (e.g., Büschken and Allenby, 2016; Büschken and Allenby, 2020) lack of the topic interpretability. In these studies, they report the estimated interpretable topics, but their models do not include any function to ensure the topic interpretability, and in fact, we often faced with the estimated topics with little semantic coherence in practice data analysis. In Section 4.4, this study proposed model that fix some representative words for each product attribute to the corresponding topic to consequently make the estimated topics interpretable. Therefore, companies and managers can obtain suggestive customer feedback through customer review analysis with ensured topic interpretability.

Compared to the existing model-based approaches, our model requires domain-specific expertise for the analyst to specify representative words for certain product attributes. However, this is not a disadvantage. Commonly, companies that seek to understand consumer preference through review analysis should have some expertise in the focal domain, and the proposed model merely makes effective use of that knowledge to ensure the stability of the analysis. Also, our model is similar to the rule-based approach because we need specify correspondences between some words and product attributes in advance. But compared to creating huge rules for almost all words and phrases, the additional cost of creating rules for a few words per attribute is negligible and does not require too much advanced expertise. In other words, it can be said that this study proposes a hybrid model-based and rule-based approach.

Last contribution is that our model helps us to understand the effects of product attributes mentioned in the review text on perceived helpfulness of the review readers as well as overall satisfaction of the review writers. Modern online retailer sites providing customer reviews (e.g., Amazon and Yelp) often disclose the usefulness of the reviews received by the readers of the reviews, who can be potential future customers, in addition to overall satisfaction with the product by the past purchasing customers. By examining the relationship between product attributes mentioned in the review text and customer satisfactions, we can find out what customers are satisfied or unsatisfied with, but we can also find out what expectations potential customers have from the relationship between these attributes and helpfulness. For managers of product development / improvement, it is of course important to look into the preference structure of existing customers, but understanding the expectations formed by potential customers is also another important perspective for the long-term growth of a brand.

4.6 Conclusion

We introduced a partially labeled supervised LDA model that combines word labeling to extract interpretable perceived topics with supervised learning to explore the structure of satisfaction of experienced customers and the expectations of review readers. To obtain stable and interpretable latent topics in online customer reviews, a priori labeled words related to product attributes were assigned to respective topics. Accordingly, we connected perceived product attributes to customer satisfaction as feedback from past customers and consumer interest in products as the perceived helpfulness of future customers through supervised learning. Referring to models for satisfaction scores by review writers and perceived helpfulness by review readers, we constructed an integrated model by sequentially connecting the former model to the latter.

The model comparison demonstrates that the proposed model is reliable not only in explaining variations in customer satisfaction and reader helpfulness but also in guaranteeing interpretable and manageable topics to explain objective variables. We showed that the difference in model fit and predictive measure from the supervised LDA model is not so large and that the words obtained in the model are easy to interpret because they consist of labeled words and other words assigned to the topic according to the labeled words. In contrast, the supervised LDA model without such labeling restrictions extracts a set of words that is hardly interpretable and manageable for marketers. The model performs better in the sense of guaranteeing interpretability at the cost of model fit, which is not so significant.

In the empirical analysis, we found that the “flavor and taste,” “packing,” and “ingredient” topics were mentioned by dissatisfied customers and reviews including the “ingredient” topic were likely recognized as unhelpful by review readers. In contrast, the “health” topic had a positive effect on both customer satisfaction and perceived helpfulness. These findings open the possibility of firms and marketers controlling the level of customer satisfaction by using manageable perceived topics and identifying the attributes that customer review readers expect to find as useful to their possible future purchases.

Several problems remain. We can extend this analysis by incorporating consumer heterogeneity (Xiao, Wei, and Dong, 2016) allowing for individually different scale usage and response (Rossi, Allenby, and McCulloch, 2005) and word sentiment (Decker and Trusov, 2010; Archak, Ghose, and Ipeirotis, 2011) to accommodate the asymmetrical effects of product attributes that are positively and negatively mentioned in the review. Another aspect is predicting more “supervised” objectives such as detecting fake or deceptive reviews (Qi et al., 2016), which are meaningless, and high score ratings after false purchasing behavior. We can also extend the analysis by applying it to multiple categories. We leave these issues to future research.

Chapter 5

A Model for Customer Review Analysis by Combining Word Embedding and Topic Modeling Approach

5.1 Introduction

With the development of e-commerce sites, it has become commonplace that consumers purchase products online and give feedback on their evaluations and experiences with the products in the form of customer reviews. Companies use this wealth of information to understand consumer preference structures and make use of it in a variety of marketing activities, such as product development, market analysis, and advertising strategy planning. Therefore, developing the technology of customer review analysis plays a vital role in the modern marketing research.

In the literature, modeling the consumer behavior of creating customer reviews has been studied by many researchers to reveal the preference structure behind them. Some of them adopt the topic modeling approach, or latent Dirichlet allocation (LDA, Blei, Ng, and Jordan, 2003), to model the review generating behavior (e.g., Tirunillai and Tellis, 2014). This model assumes the existence of latent topics behind the words in the document, and these studies apply this model to the review analysis by assuming the existence of product attributes (e.g., price) behind the words in the customer review (e.g., expensive and cheap). Furthermore, the

LDA not only naturally incorporates the process of consumers recalling the product attributes into the modeling of review text generation, but also it can be easily extended to models aimed at understanding the relationship between the product attributes mentioned by the customers in the review and their satisfaction with the product (or review scores), that is, the preference structure, due to the development of topic modeling approach, such as supervised topic model (Blei and McAuliffe, 2007).

From the perspective of text modeling, however, LDA has a major problem of ignoring the ordering of words, or the context, because it treats the text data as *bag-of-words* in the word generation process. This means that a text (such as a sentence or document) is represented as the multiset of its words, while keeping multiplicity of words but disregarding grammar and even word order. Therefore, if a review describes both good and bad points of different attributes (e.g., "*This smartphone has a bright and sharp screen, but too weighty.*), the LDA regards that the words representing the good attribute (*screen*) co-occur with some words used to describe the bad attribute (*weight*), and then the LDA cannot correctly capture the relationship between the words and attributes.

The word embedding model, or word2vec (Mikolov et al., 2013), is a machine learning method that has great success in the field of text modeling. The word2vec defines the probability of word generation given the surrounding words (i.e., the skip-gram model) while it projects words into a feature space. Therefore, the word2vec can understand the word context in terms of considering the words in the window, and the word2vec regards that words related to a product attribute co-occur only with their surrounding words, which can be considered to qualify the words, not with words related to another attribute at a distance in the same document. This approach can be applied in a variety of domains, such as sentiment classification (Zhang et al., 2015) and item recommendations (Caselles-Dupré, Lesaint, and Royo-Letelier, 2018).

However, because we aim to understand preference structures behind review-generating behavior, there are not many advantages of using word2vec as it is. The feature vectors of words resulting from embedding learning are usually very high-dimensional. Moreover, each dimension cannot be interpreted as in factor analysis or principal component analysis. Therefore, even if word2vec is applied to customer review analysis, we may not know consumers' expression about specific attributes in their reviews. In this study, we propose a model for customer review analysis based on word2vec and LDA by leaning vectors with respect to not only words but also topics projected into the same feature space. The purpose of this study is to clarify the effects of product attributes mentioned in customer reviews on the customers' satisfactions while considering the contexts by combining the word embedding model and topic model.

The combination of the topic model and word2vec itself has been proposed in Moody (2016)'s LDA2vec, however, this study extends his model from the following two perspectives. The first perspective is that the proposed model combines with the supervised topic model for explaining the effects of product attributes on the customer satisfactions, rather than unsupervised learning such as LDA. We can extract topics, or product attributes, in the reviews considering not only the text structure but also the relationships between the topics and the customer satisfaction through the supervised learning process. The second perspective is that this study considers not only the word embedded feature space and latent topics but also the polarity of the documents in the model of word generation process. In the literature, a number of studies have different impacts on satisfaction between positively and negatively mentioned product attributes (e.g., Decker and Trusov, 2010), which is based on the findings that consumers generally have different preference structures for overall satisfaction when they are satisfied and unsatisfied with individual product attributes (e.g., Kano model, Kano et al., 1984).

The rest of this chapter is organized as follows. Section 5.2 discusses related works in the relevant body of literature. Next, Section 5.3 describes the model structure and its estimation procedure. The empirical study in Section 5.4 and 5.5 apply the proposed model to a real dataset on e-commerce sites about cosmetics to demonstrate how the proposed model holds advantages over comparative models and what findings it provides. Finally, Section 5.6 provides the concluding remarks and directions for future research.

5.2 Literature Review

In the literature, researchers proposed some approaches using interviews and questionnaires to clarify the consumer preference structure (Fischer et al., 1999; Hoeffler, 2003). However, because these approaches are costly to implement and the obtained data are limited, using new data sources is necessary, such as information on the Internet (Netzer et al., 2008). As alternatives to the approaches using interviews and questionnaires, the customer review analysis has been studied in the marketing literature, and in this section, we review the existing studies on the proposed method for customer review analysis to clarity the contributions and positioning of this study.

Customer review analysis consists of two processes: the extraction of the product attributes from the review text and the estimation of the preference structure using econometric models, and first, we summarize the former one. To extract product attributes from review texts, some studies partially or mainly use latent variable models, such as latent Dirichlet allocation (LDA) and conditional random fields (CRF), that assume the latent attributes behind the words in the review text. However, conventional LDA and CRF cannot perform attribute extraction with sufficient accuracy, and some studies require additional human tasks. For example, Qi et al. (2016) use the LDA model and Page Rank algorithm to extract the product attributes from the text and narrow down the candidate attributes, and then they ask experts of the focal domain to categorize the candidate attributes into several product attributes. Bi et al. (2019) also extract product attributes from the text by using the LDA model and then manually integrate these attributes with similar meanings to improve the quality of the estimated attributes.

After estimating which product attributes the review author mentions in the text, they build econometric models with the attributes as explanatory variables to estimate the consumers' preference structure. Some studies build models that take into account an ordinal scale of the review ratings, such as the ordered probit models (e.g., Xiao, Wei, and Dong, 2016; Büschken and Allenby, 2016), or their discrete scale, such as the Poisson regression (e.g., Decker and Trusov, 2010), while others build linear models to explain sales (Archak, Ghose, and Ipeirotis, 2011) or reconstructed continuous values (Qi et al., 2016). In addition to the mentioned product attributes, they consider the valence of the attributes, like positive and negative (Decker and Trusov, 2010; Qi et al., 2016; Xiao, Wei, and Dong, 2016) or more than three levels (Bi et al., 2019, and this study). As discussed in the previous section, the consideration of attribute valence is based on the findings that consumers generally have different preference structures for overall satisfaction when they are satisfied and unsatisfied with individual product attributes (e.g., Kano model, Kano et al., 1984). Also, Xiao, Wei, and Dong (2016) and this study consider the brand heterogeneity into the preference model by introducing heterogeneous coefficients to capture the varying sensitivity of consumers for each brand.

Furthermore, a remarkable feature of this study is that we consider the effects of consumer attributes such as age and reviewer status. In the proposed models, we introduce two different effects of consumer attributes on the satisfaction structure: (i) the direct effect on review ratings when consumer attributes are added into the preference structure model as explanatory variables, and (ii) the indirect effect through the effects on the importance of product attributes for the review writers, considering the extended LDA with hierarchical structure for capturing the indirect effect. These make it possible to distinguish between the direct effect of product attributes and the general trends based on consumer attributes on the preference structure and to understand the importance of products attributes varying for each consumer. Thus, the proposed models can provide more valuable results for marketing activities than the existing approaches can.

Some studies take these two processes sequentially (Decker and Trusov, 2010; Qi et al., 2016; Xiao, Wei, and Dong, 2016), while others estimate a single model by integrating them (**Buschken2017**; Büschken and Allenby, 2016). The advantage of the sequential approach is that the additional human tasks, as described above, can improve the accuracy of attributes extraction by the statistical models such as LDA and CRF, while that of the integrated approach is that the fitting to the data is generally better than the sequential approach because it allows the integrated model to find the dimensions of product attributes while taking into account their impacts on the preference structure. While the integrated approach has such advantages, it does not take the additional tasks as the sequential approach, and therefore we address some issues of simple text models such as the LDA. One of the issues is that the model does not take into account the word order and grammar because it takes a unit of text data (such as sentence and document) as the bag-of-words. Büschken and Allenby (2016) tackles the problem by extending the conventional LDA model to assign topics to not a word but a sentence, that is n-gram model (here, n is the number of words in a sentence). They do not consider the word order but the joint distribution of words consisting of a sentence, and can relax the bag-of-words limitation of the conventional LDA model. Also, **Buschken2017** propose the auto-correlated topics LDA model that allows the latent topic of the current word to carry over the next word topic, that is, it considers the bi-gram model.

In this study, we propose a word embedding model with consideration of word topics to extract product attributes from the review text while relaxing the bag-of-words limitations. The word embedding model, or word2vec (Mikolov et al., 2013), was proposed in the field of natural language processing and has attracted attention because of its extraordinary performance in many natural language tasks, such as text summarization and translation, and are still being actively studied while many extended models have been proposed. Compared with the existing text models, the standout feature of the word2vec model is treating words as dense vectors, not sparse vectors such as one-hot encoding, by embedding them in a feature space and modeling the word generation process while taking into account the word order by the skip-gram model. As for the former, it flexibly incorporates the meanings of words into the statistical model by defining the word generation probability while representing a single words as a vector with hundreds of dimensions. For the generation probability of a word, the skip-gram considers the conditional probability given the surrounding words of the focal word. In predicting appropriate word in a sentence, for example, "I ... a student." (... is a masked word for prediction), the skip-gram defines the conditional probability, $p(\ldots \mid I, a, student)$. Therefore, it is expected that $p(am \mid I, a, student)$ is larger than $p(are \mid I, a, student)$ if the model learns good embedding representations.

Another limitation of the conventional LDA model is the interpretability of the estimated topics. Previous studies using the LDA have reported topics that can be interpreted by the semantic coherence of the words assigned to the topics. However, the LDA itself does not have a mechanism to guarantee the interpretability of topics, and in practice, we are often faced with some situations in which we are unable to interpret the meaning of topics even if the model appropriately estimates the topics. Such issues have not been given much consideration in the field of customer review analysis, but in the field of natural language processing, some studies have examined the evaluation of the interpretability performance of the LDA model by the coherence measure, as discussed by Mimno et al. (2011).

In this study, we use the pre-trained vectors by large corpus for the initial values of word vectors in training the proposed models to improve the interpretability of the LDA model. In the field of natural language processing, trained word embedding models using large corpus data as like Wikipedia, such as GloVe (Pennington, Socher, and Manning, 2014) and BERT (Devlin et al., 2019), have been published. Moreover, saving the training costs by giving such pre-trained vectors to the model as initial values is a common approach. Although the use of pre-trained models is, in general, considered to contribute to improved performance and faster estimation for the current tasks, this study expects it to also contribute to improved topic interpretability. This is because it is believed that the initialization of topic assignments by considering vector representations of the general meaning and use of the words will allow us to appropriately and quickly optimize the model, such that it encompasses the unique meaning to the focal domain, compared with randomly initializing the assignments in the absence of any prior information.

Finally, Table 5.1 summarizes the above discussion on the comparison of this study with the existing studies from several viewpoints of the customer review analysis.

TABLE 5.1: Comparison with existing studies

Papers	Factors on overall satisfaction				Model for preference estimation	Method of attribute extraction	Whether the order of words is considered	Preference estimation and attribute extraction
	Product attributes	Valence of attributes	Consumer attributes	Others				
Decker and Trusov (2010)	✓	Positive and Negative	-	-	Poisson regression, Negative binomial regression	Association rule, Conditional random field with hand crafted lexicon	Considered (bi-gram model)	Separated
Archak, Ghose, and Ipeirotis (2011)	✓	Levels of attributes vary for its evaluations	-	Price, Review statistics	Log linear regression	Clustering similar words using WordNet, Dependency analysis	Considered (dependency between attributes and valence)	Separated
Qi et al. (2016)	✓	Positive and Negative	-	-	Linear regression	LDA, Page rank, Hand crafted lexicon	Ignored (uni-gram model)	Separated
Xiao, Wei, and Dong (2016)	✓	Positive and Negative	-	Brand intercept	Ordered probit regression with heteroskedasticity and heterogeneity	Manually transformation, automate similarity calculation	Ignored (uni-gram model)	Separated
Büschken and Allenby (2016)	✓	-	-	-	Ordered probit regression	Sentence constrained LDA	Considered (n-gram model)	Integrated
Buschken2017	✓	-	-	-	Ordered probit regression	Auto-correlated topics LDA	Considered (bi-gram model)	Integrated
Bi et al. (2019)	✓	5 levels of valence	-	-	Ensemble neural network	LDA, Manually integration of topics	Ignored (uni-gram model)	Separated
This study	✓	Positive, Neutral and Negative	✓	Brand heterogeneity	Ordered probit regression	Word embedding model considering word topics and sentiments	Considered (skipgram-gram model)	Integrated

5.3 Model

This section introduces the proposed models, first the word embedding part and then the regression model part for the preference measurement.

5.3.1 Word Embedding Model Considering Text Topics and Sentiments

The word embedding approaches (e.g., word2vec) model the text generation process by representing words as embedding vectors into hundreds of dimensional feature space and considering their co-occurrence probability with the surrounding words. However, this study also considers topic vectors as well as word vectors, following the LDA2vec model of Moody (2016). LDA2vec constructs the text model while taking into account the context of the sentence in which the words are used by introducing the context vectors that add the word vectors to the topic vectors, which the former represents the original meaning of the words and the latter represents what topics the words are used in the sentence.

Let $\overrightarrow{w}_i = (w_{i1}, \ldots, w_{iM})^\top$ and $\overrightarrow{t}_k = (t_{k1}, \ldots, t_{kM})^\top$ be the word vector of the word i and topic vector of the topic k with a fixed embedding dimension M, respectively, where $w_{im}, t_{km} \in \mathbb{R} \forall i, k, m$. That is, words and topics are projected into the same feature space. In addition, similar to the LDA2vec, it is assumed that a topic is assigned to the word i according to the topic proportion vector of the document to which the word i belongs, $z_i \sim categorical(\theta_{d_i})$, where z_i represents the topic assignment for the word i and θ_{d_i} is the topic proportion vector which is defined below.

One of the extensions of this study from the LDA2vec model is to take into account the polarity of the documents. The proposed model reflects the sentiment proportion (negative, neutral, and positive) into the text generation model by sentiment analysis for each document. First, we apply Gilbert and Hutto (2014)'s VADER algorithm[1] of sentiment analysis for each document to obtain the sentiment proportion, $\pi_d = (\pi_{d1}, \pi_{d2}, \pi_{d3})^\top$, where π_{dl} indicates one of the three sentiment polarities, and $\sum_l \pi_{dl} = 1$. Next, we construct topic distributions that take into account the sentiment of the document by the topic proportions of each polarity $\tilde{\theta}_{dl} = (\tilde{\theta}_{dl1}, \ldots, \tilde{\theta}_{dlK_l})^\top$ weighted by the sentiment proportion, $\theta_d = (\pi_{d1} \times \tilde{\theta}_{d1}, \pi_{d2} \times \tilde{\theta}_{d2}, \pi_{d3} \times$

[1] VADER algorithm is implemented in Python's *nltk* module.

$\tilde{\theta}_{d3})^\top$, where $\tilde{\theta}_{dl}$ is assumed to follow the Dirichlet distribution as prior, $\tilde{\theta}_{dl} \sim Dirichlet(\tilde{\theta}_0)$. Therefore, the number of dimensions of θ_d, or the total number of topics, is $K_1 + K_2 + K_3$ which are the number of negative, neutral, and positive topics, respectively. In this study, unlike LDA2vec, we take topic assignments following the topic distribution that takes into account the polarity of the document.

The proposed model uses the word vectors, topic vectors, and the topic assignments to the words to construct the context vectors representing what the word means in the context, which the sum of the word vector and the topic vector corresponding the topic assignment to the word i, $\overrightarrow{c}_i = \overrightarrow{w}_i + \overrightarrow{t}_k$, if $z_i = k$. This formulation stems from the idea of natural language processing that the meaning of a word is expressed on the basis of the original meaning of the word, but it can fluctuate by taking into account the context of the document or sentence in which the word is used. The formulation of the context vectors of this study is almost same as that of Moody (2016)'s LDA2vec.

In the process of text generation model, the proposed model considers its surrounding words to define the probability of generating the focal word. We define S_i as a multiset that contains the surrounding words of the word i in the dataset. Let $I_P = \{(i,j) \mid j \in S_i\}$ and $I_N = \{(i,j) \mid j \notin S_i\}$ be the positive and negative multisets, respectively, and $I_D = I_P \cup I_N$ be the multiset of total vocabulary. Then, we define $G = \{g_{ij} \mid (i,j) \in I_G\}$, where g_{ij} is a random variable whose value is taken to be 1 if $(i,j) \in I_P$ or -1 if $(i,j) \in I_N$. In the proposed model, this random variable indicating whether the word i and the word j are in the same window is assumed to follow the binomial logit model, which the probability of the variable is defined as the standard sigmoid (or logistic) function of the inner product of the context vector and the surrounding word vector, $p(g_{ij} \mid \overrightarrow{w}_i, \overrightarrow{w}_j, \{\overrightarrow{t}_k\}, z_i) = \sigma(g_{ij} \cdot \overrightarrow{c}_i^\top \overrightarrow{w}_j)$, where $\sigma(x) = 1/1 + \exp(-x)$. Because the dependent variable of the logit model ($\overrightarrow{c}_i$) is also latent variable, this formulation can be seen as the factor model by regarding $\overrightarrow{c}_i$ as factor scores of the word i and $\overrightarrow{w}_j$ as factor loadings of the word j. However, this factor model has a constraint that factor score $\overrightarrow{c}_i$ can be decomposed into the word vector and the corresponding topic vector, $\overrightarrow{c}_i = \overrightarrow{w}_i + \overrightarrow{t}_k$.

Therefore, the likelihood of the word embedding part is provided as follows:

$$p(G, \{\overrightarrow{w}_i\}, \{\overrightarrow{t}_k\}, Z \mid \tilde{\theta}, \pi) = \prod_{i=1}^{V} \left\{ p(z_i \mid \tilde{\theta}_{d_i}, \pi_{d_i}) \prod_{j=1, i \neq j}^{V} \sigma(g_{ij} \cdot \overrightarrow{c}_i^{\top} \overrightarrow{w}_j) \right\}, \tag{5.1}$$

where V is the total number of vocabulary in the corpus. In the proposed model, the embedding vectors are trained to maximize the probability of the inner product of the context vector and the surrounding word vector, so that word vectors, which their corresponding words often belong to the same window in a dataset will have similar values. Therefore, we can obtain vector representations taking into account the co-occurrence of words within the window, that is, the context, while LDA uniformly considers the co-occurrence of words in a document.

However, the computation cost of the likelihood (5.1) is high because the number of total vocabulary is usually over thousands and the total cost of the likelihood is its square. In this study, according to the approach proposed by Mikolov et al. (2013), we also apply a negative sampling technique to approximate the likelihood (5.1) with the following computable formulation:

$$p(G, \{\overrightarrow{w}_i\}, \{\overrightarrow{t}_k\}, Z \mid \tilde{\theta}, \pi) \approx \prod_{i=1}^{V} \left\{ p(z_i \mid \tilde{\theta}_{d_i}, \pi_{d_i}) \prod_{j \in S_i} \sigma(\overrightarrow{c}_i^{\top} \overrightarrow{w}_j) \times \prod_{n \sim P_n(w)}^{N} \sigma(-\overrightarrow{c}_i^{\top} \overrightarrow{w}_n) \right\}, \tag{5.2}$$

where $P_n(w)$ is the noise distribution as a free parameter, and we choose the unigram distribution raised to the $3/4$ th power according to the Mikolov et al. (2013)'s suggestion. The number of negative samples, N, is suggested to be $5 - 20$ for a small dataset and $2 - 5$ for a large dataset. In machine learning research, since they use a huge dataset containing millions and sometimes billions words for training, our dataset is relatively small. We determine 15 words for the number of negative samples.

5.3.2 Preference Measurement Models Considering Brand Heterogeneity and Consumer Attributes

Another extension of this study from Moody (2016)'s LDA2vec model is to combine the word embedding model considering text topic and sentiment as explained above and preference measurement model considering brand heterogeneity and consumer attributes in a supervised learning fashion. Recalling that the topic assignments is assumed to follow the categorical distribution of the topic proportion, $z_i \sim categorical(\theta_{d_i})$, the topic proportion θ_d represents the summary of the product attributes mentioned in the review d, and in the preference measurement model, it works as dependent variables for explaining the customer satisfaction or the review rating score of the review d.

Let y_d be the satisfaction score of the review d that can take values from 1 to R (R depends the e-commerce site platform, for example, $R = 5$ in Amazon), and we define two ordered probit models to clarify the structure of the satisfaction scores according to the two conceivable processes that consumer attributes directly or indirectly affects the satisfaction structure. First, in the direct effect model, the consumer attributes work as dependent variables for satisfaction score to capture their direct effects on the satisfaction structure as follows.

$$\begin{aligned} y_d &= r \quad \text{if } \tau_{r-1} \leq y_d^* < \tau_r \\ y_d^* &= \alpha_{b_d} + \sum_{k=1}^{K} \beta_{b_d k}\theta_{dk} + \sum_{q=1}^{Q} X_{dq}\gamma_q + \epsilon_d, \qquad \epsilon_d \sim N(0, \sigma^2), \end{aligned} \tag{5.3}$$

where K is the total number of topics including negative, neutral, and positive topics, and the thresholds $\{\tau_r\}$ work for realizing discrete satisfaction scores y_d through the latent continuous variable y_d^*, and the both sides of thresholds, τ_0 and τ_R, and the next two thresholds, τ_1 and τ_{R-1}, are set to $-\infty$ and ∞, the points of the empirical ratio of the corresponding rating scores in the normal cumulative distribution, respectively, for model identifiability. Also, b_d indicates the brand for which the review d wrote, and we introduce the brand heterogeneity in the brand intercept α and the coefficients of the topic distribution β. X_d is dummy variables of the reviewer d's categorical attributes, such as age and status of the reviewer ranking, and γ is the coefficient vector capturing the direct effects of the consumer attributes on the satisfaction structure.

Next, we construct different model from the above direct effect model to understand the effects of the consumer attributes on the product attributes that the reviewers mention in the review; that is to say, it is the indirect effect of the consumer attributes on the satisfaction structure through the hierarchical structure of the topic distribution. We assume the Dirichlet prior for the topic distribution of each polarity $\tilde{\theta}_{dl}$ in the above direct effect model, but the indirect effect model considers the prior hierarchical structure for the topic distributions as follows.

$$\begin{aligned}
\tilde{\theta}_{dl} &= \text{softmax}\left(\bar{\theta}_{dl}\right) \\
\bar{\theta}_{dl} &= \sum_{q=1}^{Q} X_{dq}\gamma_{lq} + \lambda_{dl}, \qquad \lambda_{dl} \sim MVN(0, 0.1^2 I). \qquad (5.4)
\end{aligned}$$

Then the indirect effect model does not consider the direct effects of the consumer attributes on the satisfaction structure, $y_d^* = \alpha_{b_d} + \sum_{k=1}^{K} \beta_{b_d k}\theta_{dk} + \epsilon_d$.

Therefore, the likelihood of the preference measurement direct effect model for the customer satisfaction is provided as follows:

$$\begin{aligned}
p(Y, \tau, \alpha, \beta, \gamma, \sigma \mid \tilde{\theta}, \pi, X) &= \left\{ \prod_{d=1}^{D} p(y_d \mid y_d^*, \tau) p(y_d^* \mid \tilde{\theta}_d, \pi_d, \alpha_{b_d}, \beta_{b_d}, \gamma, \sigma, X) \right\} \\
&\quad \times p(\tau, \alpha, \beta, \gamma, \sigma), \qquad (5.5)
\end{aligned}$$

and that of the indirect effect model is provided as follows:

$$\begin{aligned}
p(Y, \tau, \alpha, \beta, \gamma, \sigma \mid \tilde{\theta}, \pi, X) &= \left\{ \prod_{d=1}^{D} p(y_d \mid y_d^*, \tau) p(y_d^* \mid \tilde{\theta}_d, \pi_d, \alpha_{b_d}, \beta_{b_d}, \sigma) \right\} \\
&\quad \times p(\tau, \alpha, \beta, \gamma, \sigma), \qquad (5.6)
\end{aligned}$$

where $p(\tau, \alpha, \beta, \gamma, \sigma)$ is the joint prior distribution for the coefficients, and the definition is explained in the Appendix A.4. Under the assumption of the conditional independence of likelihood (5.1) and (5.5) or (5.6) when the topic distributions are given, the full joint likelihood of the proposed model is obtained by the product of these equations multiplied by the prior density for the topic distribution, which is the the Dirichlet prior $p(\tilde{\theta}_{dl} \mid \tilde{\theta}_0) \sim Dirichlet(\tilde{\theta}_0)$ in the direct effect model and the multivariate normal prior $p(\tilde{\theta}_{dl} \mid X_d, \gamma_l) \sim MVN(X_d^\top \gamma_l, 0.1^2 I)$.

Finally, we briefly introduce the estimation procedure for the proposed models. In estimation procedure of the proposed model, we take a hybrid approach combining the Markov Chain Monte Carlo sampling method and the gradient-based stochastic optimization using Adam (Kingma and Ba, 2015). The stochastic optimization gives the optimal point estimates at that iteration for the two embedding vectors, and then we sample from the posterior distributions using the Metropolis-Hastings algorithm for the topic distributions and the Gibbs sampling for the remaining parameters, given the point estimates for the embedding vectors updated for every iteration. The estimation procedure and the settings of analysis in the following empirical study are explained in more detail in the Appendix A.4.

5.4 Model Comparison

5.4.1 Dataset

In the empirical study, we use customer reviews on an e-commerce site Sephora[2] which primarily dedicates to cosmetics, and they were collected by the web scraping in January 2020. This dataset consists of 8,551 customer reviews on 25 brands in mascara category, which written from January 1, 2019 to December 31, 2019. Before data analysis, the reviews were preprocessed including the removal of symbols, substituting with lowercase letters, and the removal of reviews consisting of less than 10 words. As a result, the number of words used in the final dataset is 376,033, and the number of unique words is 7,853. The number of words in a review is from 10 to 287 words. The average of the number of words in a review is 44.0 words, the median is 37 words, and the standard deviation is 29.5.

The review data include not only the text but also the review rating scores representing the customer satisfactions for the brands observed on a five-point scale. The numbers of observed rating scores from 1-star to 5-star are 917, 736, 1,009, 1,751, and 4,143. These positively skewed (J-shaped) characteristics are well-known in many previous studies on customer review analysis (e.g., Xiao, Wei, and Dong, 2016).

Table 5.2 provides summary statistics of the consumer attributes on the Sephora dataset. Beauty Insider Program is Sephora's own rewarding program, where users can take *Insider* badge for free, but can also become *VIB* or *Rouge* status for some fee to receive more discounts and free samples. Beauty Rank is an activity status that shows how much users contribute on the site, and users start at *Rookie* and move up the ranks to *Rising*, *Go-Getter* and *Boss* by writing reviews and posting figures to the gallery. Also, Sephora allows review writers to clearly indicate whether the review

[2] `https://www.sephora.com`

was written after purchasing the product or receiving free sample of the product, so that review readers can identify reviewers who use the rewarding program to receive samples (*Received*) or who are Sephora's employees (*Sephora Employee*) from the just product buyers (*Not Received*). In addition, the review writers can indicate own attributes, age, eye color, hair color, hair condition, skin tone, and skin type, in detail when writing reviews.

However, these attributes data collected by web scraping cannot be used directly for the following analysis because not all review writers reveal all the attributes. Some attributes are masked in the display of the review for some unknown reasons, and such masked attributes are shown as *Not Available* in Table 5.2. Therefore, we apply the missing value imputation method to obtain a complete attribute data by assigning the esimated attribute levels to the masked attributes. We adopted MissForest[3] Stekhoven and Buhlmann (2012) method, which was recently proposed as a missing value imputation for categorical data based on the non-parametric method. Table 5.2 shows the counts and the percentage in the attribute for each attribute level before and after missing value imputation.

5.4.2 Model Comparison Using Sephora Dataset

In this section, we demonstrate the effectiveness of the proposed models, especially consideration of the text sentiment, brand heterogeneity, and the direct and indirect effects of the consumer attributes, for in-sample fit and out-sample predictive performance. The detail of the settings for the comparison analysis such as the number of iterations and the set values for hyperparameters in the Appendix A.4.

To evaluate these features of the proposed models, we consider three comparative models, first is the supervised LDA2vec model (Model 1) which extends the LDA2vec model in a supervised learning fashion, while it does not consider any features of the proposed models. Therefore, similar to LDA2vec, Model 1 assigns topics to words according to the topic distributions that do not take into account the text sentiments, $\theta_d = (\theta_{d1}, \ldots, \theta_{dK})^\top \sim Dirichlet(\theta_0)$, and does not include coefficients for brand heterogeneity or terms for consumer attributes in the preference measurement model, $y_d^* = \alpha + \sum_{k=1}^{K} \beta_k \theta_{dk} + \epsilon_d$. The second is the supervised LDA2vec model

[3] MissForest algorithm is implemented in Python's *missingpy* module.

TABLE 5.2: Summary statistics of consumer attributes

Attribute Names	Levels	Before Imputation		After Imputation	
		Count	Percentage	Count	Percentage
Beauty Insider Program	Insider	3535	41.3	3583	30.8
	Rouge	2609	30.5	2635	27.3
	VIB	2315	27.1	2333	41.9
	Not Available	92	10.8	0	0.0
Beauty Rank	Boss	6	0.1	6	0.1
	Go-Getter	14	0.2	14	0.2
	Rising Star	9	0.1	9	0.1
	Rookie	8430	98.6	8522	99.7
	Not Available	92	1.1	0	0.0
Free Product	Received	3227	37.7	3227	37.7
	Sephora Employee	71	0.8	71	0.8
	Not Received	5253	61.4	5253	61.4
Age	13 - 17	51	0.6	471	5.5
	18 - 24	183	2.1	1582	18.5
	25- 34	370	4.3	4063	47.5
	35- 44	169	2.0	1258	14.7
	45 - 54	66	0.8	391	4.6
	Over 54	142	1.7	786	9.2
	Not Available	7570	88.5	0	0.0
Eye Color	Blue	1789	20.9	1836	21.5
	Brown	3915	73.9	4125	48.2
	Gray	60	0.7	60	0.7
	Green	1281	15.0	1285	15.0
	Hazel	1242	14.5	1245	14.6
	Not Available	264	3.1	0	0.0
Hair Color	Auburn	369	4.3	369	4.3
	Black	1295	15.1	1304	15.2
	Blonde	2089	24.4	2105	24.6
	Brunette	4205	49.2	4397	51.4
	Gray	63	0.7	63	0.7
	Red	313	3.7	313	3.7
	Not Available	217	2.5	0	0.0
Hair Condition	Chemically Treated	26	0.3	5894	68.9
	Coarse	4	0.05	284	3.3
	Curly	1	0.01	150	1.8
	Dry	5	0.06	338	4.0
	Fine	6	0.07	300	3.5
	Normal	8	0.09	975	11.4
	Oily	4	0.05	519	6.1
	Straight	4	0.05	72	0.8
	Wavy	1	0.01	19	0.2
	Not Available	8492	99.3	0	0.0
Skin Tone	Dark	196	2.3	196	2.3
	Deep	271	3.2	272	3.2
	Ebony	49	0.6	49	0.6
	Fair	1898	22.2	1902	22.2
	Light	2419	28.3	2429	28.4
	Medium	1847	21.6	1992	23.3
	Olive	603	7.1	603	7.1
	Porcelain	563	6.6	563	6.6
	Tan	544	6.4	545	6.4
	Not Available	161	1.9	0	0.0
Skin Type	Combination	4449	52.0	4621	54.0
	Dry	1479	17.3	1480	17.3
	Normal	1376	16.1	1376	16.1
	Oily	1073	12.5	1074	12.6
	Not Available	174	2.0	0	0.0

with text sentiment consideration (Model 2) which extends the Model 1 to consider the text sentiments in the same way of the proposed models discussed in Section 5.3.1. Therefore, the topic distribution is changed from the form of the Model 1 to that of the topic proportions for each sentiment weighted by the sentiment proportion of the review, but the preference measurement model still does not consider brand heterogeneity and consumer attributes. The third model extends the Model 2 to consider brand heterogeneity in the preference measurement model, that is, $y_d^* = \alpha_{b_d} + \sum_{k=1}^{K} \beta_{b_d k}\theta_{dk} + \epsilon_d$, and we call it Model 3 in the following. A total of five models, these comparative models and the proposed models, the direct effect model (Model 4) and the indirect effect model (Model 5), are used to compare the fitting performance for 90% of the available data and the predictive performance for the remaining 10% out-sample data.

To evaluate the performance for the above models, we apply four measures with different perspectives: the log marginal likelihood (LMD) for in-sample data and out-sample data calculated based on the harmonic mean of the log likelihood (Newton and Raftery, 1994), the widely applicable information criterion (WAIC, Watanabe, 2010) using the in-sample log likelihood to evaluate the generalization error, and the mean squared error (MSE) to calculate the predictive accuracy for the ordered categorical review ratings in the out of samples.

Figure 5.1 reports the calculated values for each measure and model in the range of the varying number of topics from 1 to 10 (since these numbers indicate the number of topics for each polarity, the total number of topics is three times of those numbers). When compared by the LMD measure for in-sample and out-sample data, almost models compete each other by the number of topics, but for the best values (the number of topics is 8 and 9), Model 3 outperforms the others. WAIC comparisons similarly show that Model 3 outperform the others, but the proposed direct effect model (Model 4) also competes it. However, the proposed indirect effect model (Model 5) is worse than others in the almost measures' comparisons. Thus, improving the proposed models, including the revision of the structure of the indirect effect model and the addition of new variables and structures to improve the predictive accuracy for review ratings, is left for the future work.

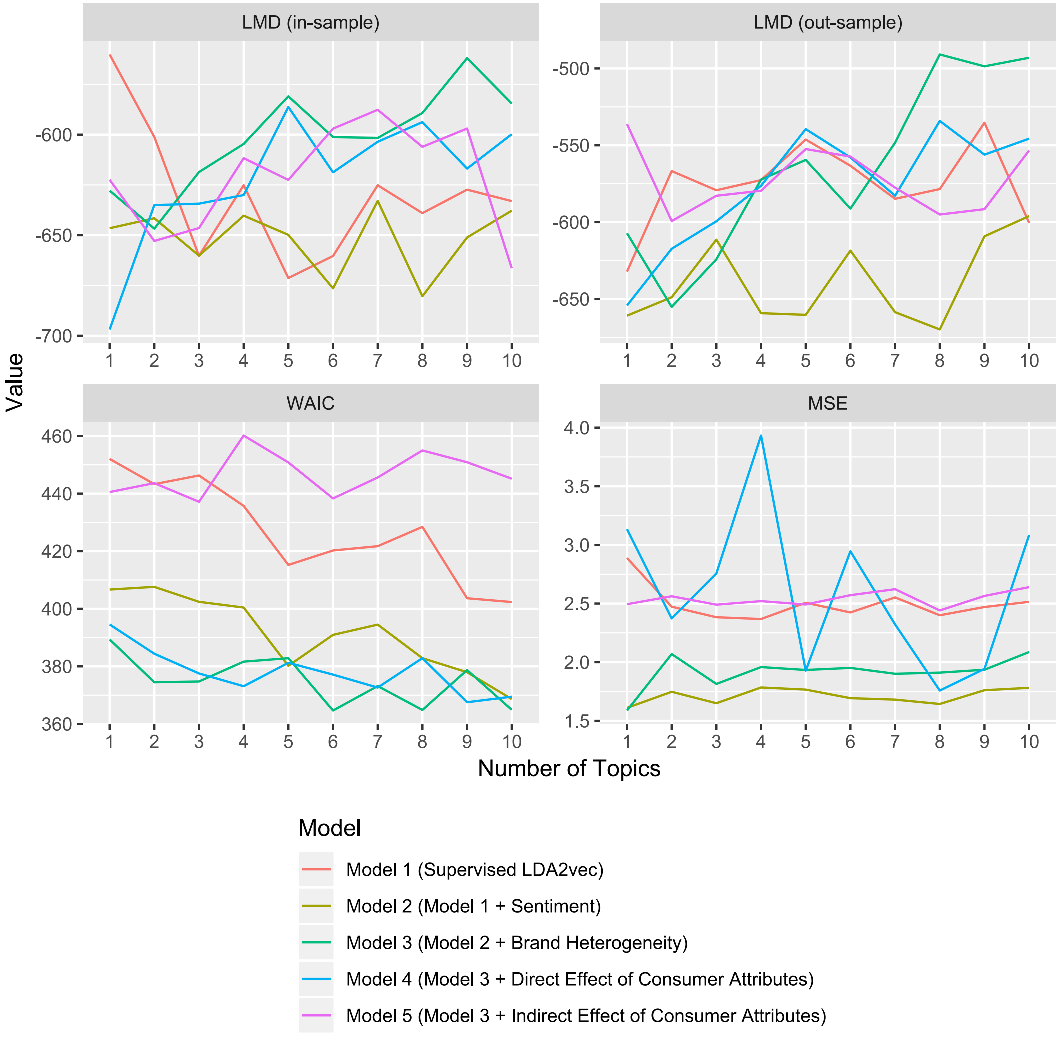


FIGURE 5.1: Values of LMD (in-sample), LMD (out-sample), WAIC, and MSE for model comparison

In the next section, we report the estimation results of the proposed direct effect model with eight topics for each polarity which is the best value in the almost comparisons.

5.5 Discussion

5.5.1 Empirical Results

First, we interpret each dimension of the estimated topics representing the product attributes mentioned in the review text by the semantic coherence of the meanings of the words associated with the topics. In the proposed word embedding model, since we obtain the embedding vectors for words and topics on the same feature space, we can consider the words whose vectors are closest to the topic vector as the most related words to that topics.

Table 5.3 displays the top 10 words corresponding to the closest word vectors to each topic vector. But because the estimated model has eight topics for each polarity, the table shows two topics for each polarity for saving the space. These words for each topic provide interpretable semantic coherent about mascara products, for example, negative topic 2 discusses complaints about mascara peeling off easily as some words suggest, such as *horribly*, *smeared*, and *flakes*. Topic 5 discussed the smell emitted by the mascara products, which cannot necessarily be judged as negative opinions from the words in the table (such as *smell*, *fragrance*, and *perfumy*). However, this topic is often in the negative reviews and thus deemed to be a discussion of unpleasant odors. Both topics 10 and 11 discuss the functionality of the mascara products, which relate to their effects on eyelashes and the resistance of the mascara to peel off, respectively. The product reviews in the dataset include topics on not only the performance of the products and impressions after use, but also the packaging design of the products (topic 18) and declarations that free samples were provided to the writers for writing reviews (topic 19). It was estimated that these topics tended to be mentioned along with positive reviews.

Next, we discuss the estimation results of the preference model, and Figure 5.2 shows the estimates of the intercept and coefficient of the topic distributions considering brand heterogeneity. Since the topic distributions of the proposed model take into account the review writers' sentiments inferred from the review text (not the review ratings), the estimates capture an overall positive correlation between the sentiment and ratings for each polarity (topics 1 to 8 are negative topics, topics 9 to 16 are neutral topics, and the remaining topics are positive topics). Furthermore, these parameters are heterogeneous with respect to brands, and they capture the impacts of the proportion of the estimated product attributes discussed in the review, which vary across brands, on review ratings. For example, brands 1, 18, and 24 have large positive impacts on the positive topics, but they also have large negative impacts on the negative topics, indicating that the brand satisfaction fluctuate greatly, depending on the customers' sentiments about their perceived product attributes. On the other hand, brands 12 and 19 have relatively small absolute values for both positive and negative topics, indicating a weak correlation between the customer's sentiment for the product attributes of these brands and the satisfaction with the products.

Finally, Table 5.4 provides the posterior means of the coefficients for consumer attributes, the thresholds, and variance parameter. The table also provides the 90% highest posterior density (HPD) next to posterior means, and the bold face indicates that the estimates are significantly far from zero. The estimates of the threshold parameters (τ_r) indicate that an approximate 0.4 increase in the latent continuous rating (y_d^*) is associated with a one-point increase in the observed discrete rating (y_d). For example, if the topic proportion for topic 23 in the review for the products of the brand 1 increase by 10%, the expected change in the latent rating is 0.39, translating to an almost one-point increase in customer satisfaction.

The estimates of the coefficients of consumer attributes also show some interesting findings, for example, the effect of *VIB* rank, who pay the highest prices in the rewarding program, negatively estimated, indicating these customers, on average, less satisfied in the mascara category. Also, those who did not receive free product samples were 0.19 less satisfied on the latent continuous rating scale, which is translated as almost 0.5 point decrease on the observed discrete rating scale. This result also corresponds to the above result that topic 19 on free samples was estimated to be associated with positive reviews. Other estimated results also show some significant relationships between reviewers' statement about their own attributes and the satisfaction scores, for example, the effects of the middle age (25 to 34), gray eye color, and normal and oily hair condition are negatively estimated, while the effect of gray hair color is largely positively estimated.

TABLE 5.3: Top 10 words whose vectors are closest to the topic vectors of topic 2, 5, 10, 11, 18, and 19

Topic 2 (Negative)	Topic 5 (Negative)	Topic 10 (Neutral)
Flaking of Mascara	Smell	Performance on Eyelash
horribly	smell	lashes
odour	fragrance	look
smeared	blepharitis	curled
flecked	perfumy	makes
flakes	nasty	long
droplet	similiarly	them
gross	bore	straight
smears	posters	and
under	seemingly	eyelashes
terribly	chemical	longer
Topic 11 (Neutral)	Topic 18 (Positive)	Topic 19 (Positive)
Easy to peel off	Free Samples	Packaging
off	influenster	unique
day	von	cute
wash	free	nice
face	from	packaging
hours	kat	fanning
water	test	color
panda	received	sleek
pool	beauty	black
remover	cruelty	separates
under	sent	brush

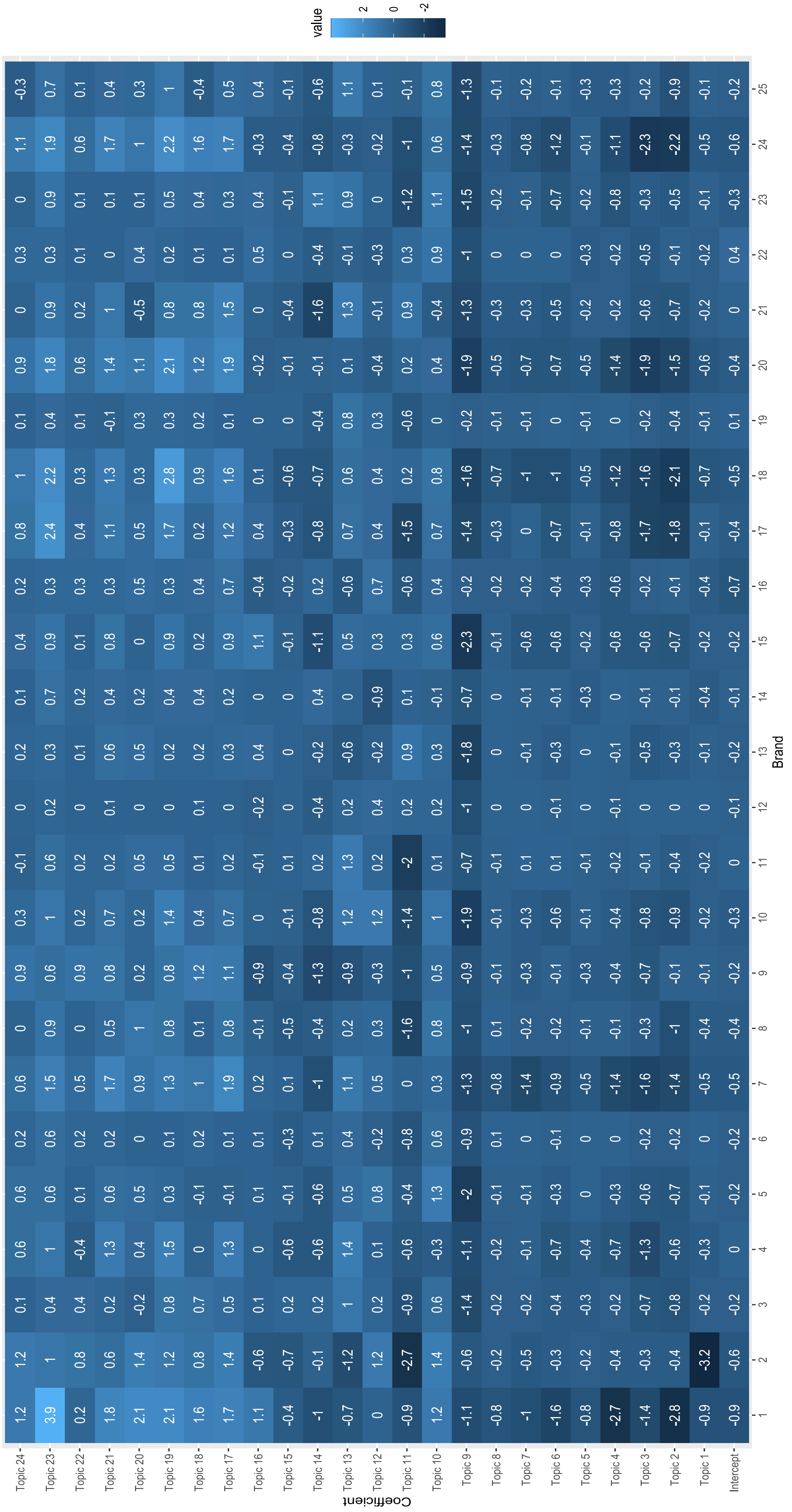


FIGURE 5.2: Estimates of brand intercept and brand heterogeneous coefficients of topic distribution

TABLE 5.4: The estimates of the coefficients for consumer attributes, thresholds, and variance parameter

Variables	Levels	Posterior Mean	HPD Interval (90%)		Variables	Levels	Posterior Mean	HPD Interval (90%)	
			Min	Max				Min	Max
Beauty Insider Program	Rouge	-0.02	-0.08	0.04	Hair Condition	Coarse	-0.05	-0.18	0.08
	VIB	**−0.07**	-0.13	-0.01		Curly	0.01	-0.17	0.20
Beauty Rank	Go-Getter	0.01	-0.56	0.61		Dry	0.02	-0.11	0.15
	Rising Star	0.55	-0.15	1.25		Fine	-0.04	-0.17	0.10
	Rookie	0.10	-0.26	0.47		Normal	**−0.09**	-0.17	-0.01
Free Product	Sephora Employee	0.08	-0.15	0.33		Oily	**−0.14**	-0.26	-0.02
	Not Received	**−0.19**	-0.25	-0.12		Straight	-0.15	-0.41	0.12
Age	18 - 24	-0.09	-0.21	0.03		Wavy	-0.35	-0.82	0.11
	25 - 34	**−0.13**	-0.24	-0.01	Skin Tone	Deep	0.09	-0.09	0.28
	35 - 44	-0.07	-0.20	0.07		Ebony	0.03	-0.29	0.34
	45 - 54	-0.18	-0.34	-0.03		Fair	-0.03	-0.17	0.14
	Over 54	-0.04	-0.17	0.11		Light	-0.01	-0.17	0.14
Eye Color	Brown	0.04	-0.03	0.11		Medium	-0.11	-0.27	0.04
	Gray	**−0.27**	-0.55	-0.02		Olive	0.06	-0.11	0.23
	Green	-0.02	-0.10	0.05		Porcelain	0.03	-0.14	0.21
	Hazel	0.03	-0.05	0.11		Tan	0.07	-0.11	0.24
Hair Color	Black	0.02	-0.10	0.15	Skin Type	Dry	0.01	-0.06	0.07
	Blonde	0.11	-0.02	0.22		Normal	-0.01	-0.08	0.05
	Brunette	0.08	-0.04	0.19		Oily	0.03	-0.04	0.10
	Gray	**0.52**	0.21	0.85					
	Red	0.11	-0.03	0.28					
Thresholds	Rating 1-2 (τ_1)	-1.88	Fixed			Rating 3 - 4 (τ_3)	**−0.96**	-0.97	-0.95
	Rating 2 - 3 (τ_2)	**−1.42**	-1.44	-1.41		Rating 4 - 5 (τ_4)	-0.31	Fixed	
Variance	σ^2	0.91	0.87	0.95					

5.5.2 Marketing Values of This Study

Contribution of this study on the academic literature is that we construct new marketing models for customer review analysis combining word embedding model and supervised sentiment topic model, but marketing values of the proposed model are given by incorporating consumer attributes into the model structure. In the marketing literature, as discussed in Section 5.2, previous models for customer review analysis do not take into account the effects of the consumer attributes on the overall satisfactions (direct effects) and the proportions of product attributes mentioned in the review text (indirect effects).

However, our models incorporating these direct or indirect effects can be valuable for practitioners as well as academics. For example, the direct model revealed that several consumer attributes significantly affect overall satisfaction, such as free sample availability and user status of the rewarding program, as shown in the previous section. The impacts of product attributes mentioned in the review text is the estimates excluding the impact of these consumer attributes, and if we ignored these consumer attributes, the estimated results might lead to the selection of wrong marketing activities due to the omitted variable bias. Our model is also valuable to the review platform managers because it allows them to add their unique values to the product introduction in the platform by using the impacts of consumer attributes, such as "this product is highly rated by women teenagers / people who have trouble with dry hair." At first glance, this seems feasible through simple descriptive statistics, but since it does not take into account the effect of product attributes in the text, such analysis is inappropriate due to the omitted variable bias as well.

Although the results of the indirect effect model are not shown in the previous section, it also can provide some effective implications. Indirect model allows us to understand what extent the commonality by consumer attributes affects the variations of the product attribute proportions for each review. This idea was originally developed in the context of the hierarchical Bayesian models (e.g., Rossi, Allenby, and McCulloch, 2005), and this study follows them to stably estimate the model considering the variation in the customer reviews by assuming that the heterogeneity of the product attribute proportion, i.e., the topic distribution per review, is distributed around the commonality explained by consumer attributes. This formulation allows companies and managers to know which customer segments have interest in what attributes, such as women in their twenties have high interest in the attribute of ease to peel mascara because they often mention in their reviews, which can be effective implications for the product development and advertising planning.

5.6 Conclusion

We introduced a model for customer review analysis by combining the word embedding approach and the topic modeling approach. The purpose of this study is to consider the context in the review text beyond the limitations of the bag-of-words assumption, which ignores the word order assumed in the conventional topic models, and to clarify the effects of the product attributes mentioned in the reviews on the customer satisfactions for the products. The combination of the word embedding model and topic model itself has been proposed in Moody (2016)'s LDA2vec model, however, we extend his model from the two perspectives: the supervised learning and the consideration of the text sentiments. The embedding vectors of words and topics are optimized based on not only the fitting to the text data but also their impacts on the structure of the review ratings through the topic proportion parameters for each review text. Also, the word topics, which are assigned to words in considering the context of the review text, are determined by the interaction between the topic and the polarity proportions obtained through the sentiment analysis in advance.

Furthermore, the proposed model sophisticates the preference structure from two perspectives from the previous studies. First is the assumption of brand heterogeneity in the impacts of the topic proportions on the review ratings, and this extension allows us to estimate the individual effects of the sentiments for each product attribute in the review on the product overall satisfaction for each brand. Second is to consider the impact of consumer attributes on the satisfaction, and this study proposed two models that estimate the direct effect of consumer attributes by introducing attributes as dependent variables in the ordered probit model for preference measurement, and the indirect effect by introducing them into the regression model in the hierarchy structure of the topic proportions.

In the empirical analysis, we constructed some comparative models by subtracting the features of the above proposed models to demonstrate their effectiveness for a real data analysis through the model comparison using Sephora dataset. The comparison results show that the proposed direct effect model outperforms than others in the sense of the model fitting, predicting performance, and generalization error. However, the indirect effect model is inferior to others, and all of these models have the poor accuracy of predicting review rating in the out of sample data, and these issues are left for the future works. The estimation results using the best model in the model comparison show that the model estimates some interpretable product attributes, such as flaking, smell and eyelash performance of the mascara product, and provides some interesting findings on the preference structure, such as the heterogeneous impacts of the topic proportions in the review on the overall satisfaction for the brands and the positive and direct effect of receiving free samples on the satisfaction.

In addition to the issues above, some extensions should be addressed. First, we should also evaluate how well the proposed model explains the text data as a language model, unlike the fitting and predicting performance to the review ratings demonstrated in this study. The effects of the supervised learning and the consideration of the text sentiments, which are the main contributions on the literature of the language models and the review analysis, on understanding the language structure of the review text must be evaluated. Second, other effects of consumer attributes on the helpfulness structure of the customer review must be considered, similarly to the study of Chapter 4. Furthermore, it is more interesting to clarify the interaction effects between the levels of consumer attributes and the topic proportions, for example, reviews referring to the product attributes of the performance on eyelashes may be more useful to the review readers if the information on the review writers' eye color is displayed.

Appendix A

Estimation Procedures

A.1 Derivation of the Collapsed Gibbs Sampler for the MM-STB

In Section 1.3.2, we derived the conditional posterior distributions of latent variables (Equations (1.4) and (1.5)). To derive these posteriors, we need the full conditional posterior distributions for model parameters, and these are given as follows:

$$\begin{aligned}
&P(\eta_i|S,R,X,\gamma)\\
&= \frac{\Gamma\left(\sum_k N_{ik}+M_{ik}+\gamma_k\right)}{\prod_k \Gamma(N_{ik}+M_{ik}+\gamma_k)} \prod_{k=1}^{K} \eta_{ik}^{N_{ik}+M_{ik}+\gamma_k} && \text{(A.1)}\\
&P(\psi_{kk'}|A,S,R,\delta,\epsilon)\\
&= \frac{\Gamma(n_{kk'}^{(+)}+n_{kk'}^{(-)}+\delta_{kk'}+\epsilon_{kk'})}{\Gamma(n_{kk'}^{(+)}+\delta_{kk'})\Gamma(n_{kk'}^{(-)}+\epsilon_{kk'})} \times \psi_{kk'}^{\mathbb{I}(a_{ij}=1)}(1-\psi_{kk'})^{\mathbb{I}(a_{ij}=0)} && \text{(A.2)}
\end{aligned}$$

$$\begin{aligned}
P(\theta_k|X,Z,\alpha) &= \frac{\Gamma\left(\sum_l M_{kl}+\alpha_l\right)}{\Pi_l \Gamma(M_{kl}+\alpha_l)} \prod_{l=1}^{L} \theta_{kl}^{M_{kl}+\alpha_l} && \text{(A.3)}\\
P(\phi_l|W,Z,\beta) &= \frac{\Gamma\left(\sum_v M_{lv}+\beta_v\right)}{\Pi_v \Gamma(M_{lv}+\beta_v)} \prod_{v=1}^{V} \phi_{lv}^{M_{lv}+\beta_v}, && \text{(A.4)}
\end{aligned}$$

where N_{ik} is the count number of when node i is assigned community k on the edges from node i to other nodes and from other nodes to node i. M_{ik} is the count number of when words in node i's document are assigned to community k. $n_{kk'}^{(+)}$ $(n_{kk'}^{(-)})$ is the number of links (non-links) from nodes in community k to nodes in community k'. M_{kl} is the count number of when words are assigned to community k and topic l. M_{lv} is the count number of when word v is assigned to topic l. Γ is the gamma function, and $\mathbb{I}$ is the indicator function that returns 1, if the condition is satisfied, and 0 otherwise.

Collapsed Gibbs sampling repeats the sampling procedure according to Equations (1.4) and (1.5). The pseudo algorithm for the proposed model is provided in algorithm 1.

Algorithm 1 collapsed Gibbs sampler for MMSTB

```
Assign randomly communities and topics to $S, R, X, Z$
for $g = 1, \dots, G$ do
    for $i = 1, \dots, D$ do
        for $j = 1, \dots, D$ do
            Set $N_{ik\backslash ij}, N_{jk'\backslash ji}, n^{(+)}_{kk'\backslash ij}, n^{(-)}_{kk'\backslash ij}$
            Sample edge communities, $s^{(g)}_{ij}, r^{(g)}_{ji}$, from (1.4)
            Update $N_{ik}, N_{jk'}, n^{(+)}_{kk'}, n^{(-)}_{kk'}$
        end for
        for $m = 1, \dots, M_i$ do
            Set $M_{ik\backslash im}, M_{kl\backslash im}, M_{lv\backslash im}$
            Sample word community and word topic, $x^{(g)}_{im}, z^{(g)}_{im}$, from (1.5)
            Update $M_{ik}, M_{kl}, M_{lv}$
        end for
    end for
end for
```

A.2 Posterior Distributions of Dynamic Topic Model for Social Influence

In this appendix, we define the posterior distributions of the dynamic topic model introduced in Chapter 3, and its MCMC algorithm. First, we apply the collapsed Gibbs method for sampling the topic assignment z by integrating out element distribution ϕ. The conditional probability density of the topic assignment $z_{dtn} = k$ is given as:

$$p(z_{dn} = k) \mid \eta_{dt} \propto \frac{\exp(\eta_{dtk})}{\sum_{k'} \exp(\eta_{dtk'})} \times \frac{N_{tkv\backslash dtn} + \phi_0}{\sum_{v'} N_{tkv'\backslash dtn} + \phi_0}, \tag{A.5}$$

where N_{tkv} represents the counts of assignments of topic k into element v at time t.

Next, we derive the conditional posterior distributions of self-influences and time-specific random effects. In this study, we assume that these parameters follow the normal distribution as prior, $\alpha_{dk} \sim N(0, \sigma^2_{\alpha 0})$ and $\gamma_{tk} \sim N(0, \sigma^2_{\gamma 0})$. Then the conditional posterior distributions are given as:

$$
\begin{aligned}
p(\alpha_{dk} \mid \eta_{d \cdot k}, \beta, \gamma, \sigma_\epsilon, \sigma_{\alpha 0}) &\propto N(\mu_{\alpha dk}, \sigma^2_{\alpha dk}), \\
\sigma^2_{\alpha dk} &= \left(\frac{\sum_t \eta^2_{dt-1k}}{\sigma^2_\epsilon} + \frac{1}{\sigma^2_{\alpha 0}} \right)^{-1}, \\
\mu_{\alpha dk} &= \sigma^2_{\alpha dk} \left(\frac{\sum_t \eta_{dt-1k} \left(\eta_{dtk} - \sum_f \beta_{dfk} \eta_{ft-1k} - \gamma_{tk} \right)}{\sigma^2_\epsilon} \right),
\end{aligned} \tag{A.6}
$$

and

$$
\begin{aligned}
p(\gamma_{tk} \mid \eta_{\cdot \cdot k}, \alpha, \beta, \sigma_\epsilon, \sigma_{\gamma 0}) &\propto N(\mu_{\gamma tk}, \sigma^2_{\gamma tk}), \\
\sigma^2_{\gamma tk} &= \left(\frac{D}{\sigma^2_\epsilon} + \frac{1}{\sigma^2_{\gamma 0}} \right)^{-1}, \\
\mu_{\gamma tk} &= \sigma^2_{\gamma tk} \left(\frac{\sum_d \eta_{dtk} - \alpha_{dk} \eta_{dt-1k} - \sum_f \beta_{dfk} \eta_{ft-1k}}{\sigma^2_\epsilon} \right),
\end{aligned} \tag{A.7}
$$

Therefore, the algorithm of the MCMC procedure for the dynamic topic model is as follows:

1. initialize $\eta, Z, \zeta, \alpha, \beta, \gamma, \delta$
2. iterate the following samplers until all parameters converge
   (a) sample η using the forward filtering backward sampling scheme introduced in Section 3.4.3
   (b) sample Z according to equation (A.5)
   (c) sample ζ according to $p(\zeta_{dtk} \mid \eta_{dt\cdot}) \sim PG(N_{dt}, \psi_{dtk})$
   (d) sample α and γ, according to equations (A.6) to (A.7)
   (e) sample β and hyperparameters of shrinkage prior according to Section 3.4.2

3. calculate the expectations of the element distribution using the last samples of Z according to

$$\phi_{kv} = \frac{N_{kv} + \phi_0}{\sum_{v'} N_{kv'} + \phi_0}$$

In the empirical study, we repeat the above MCMC process 1,000 times and then use last 800 samples to calculate the posterior means and intervals of HPD. The settings of the prior distributions used in the empirical study are as follows:

$$\begin{aligned}
\phi_k &\sim Dirichlet(\phi_0), \quad \phi_0 = 0.1 \\
\alpha_{dk} &\sim N(0, \sigma^2_{\alpha 0}), \quad \sigma^2_{\alpha 0} = 100.0 \\
\gamma_{tk} &\sim N(0, \sigma^2_{\gamma 0}), \quad \sigma^2_{\gamma 0} = 100.0
\end{aligned}$$

A.3 Posterior Distributions of PLS-LDA Model

In this appendix, we describe the details of the posterior distributions of our PLS-LDA model and the MCMC algorithm. First, we apply the collapsed Gibbs method for sampling the topic assignment variable z by integrating out topic distribution θ and word distribution ϕ. The conditional probability density of topic assignment $z_{dn} = k$ is given as

$$\begin{aligned}
& p(z_{dn} = k \mid w_{dn} = v_k, W_{\backslash dn}, y^*_{s,d}, y_{h,d}, x_{s,d}, x_{h,d}, Z_{\backslash dn}, \alpha, \beta^*, \gamma_s, \gamma_h, \delta_s, \delta_h, \tau) \\
& \propto \left(N_{dk\backslash dn} + \alpha\right) \frac{N_{kv_k\backslash dn} + \beta^*_{kv_k}}{\sum_{v=1}^{V_k} N_{kv\backslash dn} + \beta^*_{kv}} p(y^*_{s,d} \mid z_{dn} = k, x_{s,d}, \gamma_s, \delta_s, \tau) \\
& \quad p(y_{h,d} \mid z_{dn} = k, x_{h,d}, y_{s,d}, \gamma_h, \delta_h), \qquad \text{(A.8)}
\end{aligned}$$

where N_{kv} represents the counts of assignments of topic k into word v and the symbol $\backslash$ represents the exclusion of the word from the counts. $p(y^*_{s,d} \mid \cdot)$ and $p(y_{h,d} \mid \cdot)$ are the probability density functions of the normal distribution and the Poisson distribution, respectively.

Next, for the ordered probit model of satisfaction scores, we apply Gibbs sampling with data augmentation. Using results from the existing literature, the conditional densities of the regression coefficients γ_s and δ_s, the augmented continuous satisfaction $y^*_{h,d}$, and the threshold parameters τ are multivariate normal, truncated normal, and uniform distribution, respectively.

$$
\begin{aligned}
& p(\gamma_s \mid Y^*_s, Z, X_s, \delta_s, g_{s,0}) \sim N(\mu_{\gamma_s}, \Sigma_{\gamma_s}), \\
& \quad \Sigma_{\gamma_s} = \left(\sum_{d=1}^{D} \log(N_{d\cdot} + 1) \log(N_{d\cdot} + 1)^\top + g^{-1}_{s,0} \cdot I \right)^{-1}, \\
& \quad \mu_{\gamma_s} = \Sigma_{\gamma_s} \left(\sum_{d=1}^{D} \log(N_{d\cdot} + 1) \left(y^*_{s,d} - \sum_{m=1}^{5} \delta_{s,m} x_{s,dm} \right) \right) && \text{(A.9)} \\
& p(\delta_s \mid Y^*_s, Z, X_s, \gamma_s, d_{s,0}) \sim N(\mu_{\delta_s}, \Sigma_{\delta_s}), \\
& \quad \Sigma_{\delta_s} = \left(\sum_{d=1}^{D} x_{s,d} x^\top_{s,d} + d^{-1}_{s,0} \cdot I \right)^{-1}, \\
& \quad \mu_{\delta_s} = \Sigma_{\delta_s} \left(\sum_{d=1}^{D} x_{s,d} \left(y^*_{s,d} - \sum_{l=1}^{L} \gamma_{s,l} \log(N_{dl} + 1) \right) \right) && \text{(A.10)} \\
& p(y^*_{s,d} \mid y_{s,d}, z_d, x_d, \gamma_s, \delta_s, \tau) \sim N \left(\sum_{l=1}^{L} \gamma_{s,l} \log(N_{dl} + 1) + \sum_{m=1}^{5} \delta_{s,m} x_{s,dm}, 1 \right), \\
& \quad \text{truncated to } (\tau_{r-1}, \tau_r] \text{ if } y_{s,d} = r && \text{(A.11)} \\
& p(\tau_r \mid Y_s, Y^*_s, \tau_q) \sim U[\tau^*_{\text{lhs}}, \tau^*_{\text{rhs}}], \quad r = 1, \ldots, R-1, q \neq r \\
& \quad \tau^*_{\text{lhs}} = \max\left(\max\{y^*_{s,d}; y_{s,d} = r\}, \tau_{r-1} \right) \\
& \quad \tau^*_{\text{rhs}} = \min\left(\min\{y^*_{s,d}; y_{s,d} = r+1\}, \tau_r \right) && \text{(A.12)}
\end{aligned}
$$

Finally, we employ the random walk Metropolis-Hastings algorithm to estimate coefficients γ_h and δ_h in Poisson regression for helpfulness. The joint conditional density of γ_h and δ_h is given by the product of the Poisson density for Y_h and the normal density for the prior distribution. Because the constant term of this posterior density is unknown and obtaining samples from the posterior is not easy, we employ the Metropolis-Hastings algorithm for sampling from the posterior. The proposal density is the normal distribution with the mean of its own values from the previous iteration, $\gamma^{(t)}_h \sim N(\gamma^{(t-1)}_h, \sigma^2_{\gamma_h} \cdot I)$. $\sigma^2_{\gamma_h}$ is a step-size parameter whose value is adjusted in the MCMC procedure so that the acceptance rate falls into the range between 30% and 50%. Because the proportion of proposal densities can be cancelled, the acceptance ratio in sampling $\gamma^{(t)}_h$ consists of the proportion of posterior distributions:

$$\begin{aligned}
\alpha_{\gamma_h} &= \min\left\{1, \frac{p(Y_h \mid Z, X_h, \gamma_h^{(t)}, \delta_h)p(\gamma_h^{(t)} \mid g_{h,0})}{p(Y_h \mid Z, X_h, \gamma_h^{(t-1)}, \delta_h)p(\gamma_h^{(t-1)} \mid g_{h,0})}\right\}, \\
p(Y_h \mid Z, X_h, \gamma_h^{(t)}, \delta_h) &= \prod_{d=1}^{D} Po(y_{h,d}; \sum_{l=1}^{L} \gamma_{h,l}^{(t)} \log(N_{dl}+1) + \sum_{m=1}^{6} \delta_{h,m} x_{h,dm}) \\
p(\gamma_h^{(t)} \mid g_{h,0}) &= N(\gamma_h^{(t)}; 0, g_{h,0}^{-1} \cdot I)
\end{aligned} \tag{A.13}$$

δ_h is also sampled in the same way.

Therefore, the algorithm of the MCMC procedure is as follows:

1. initialize $Z, \gamma_s, \delta_s, Y_s^*, \tau, \gamma_h, \delta_h, \sigma_{\gamma_h}, \sigma_{\delta_h}$
2. iterate sampling until all parameters converge
   (a) sample Z according to equation (A.8)
   (b) sample $\gamma_s, \delta_s, Y_s^*, \tau$ according to equations (A.9) to (A.12)
   (c) update γ_h and δ_h with the acceptance ratio (A.13)
   (d) adjust σ_{γ_h} and σ_{δ_h} if the cumulative number of the acceptance falls outside the desired percentage
3. calculate the expectations of θ and ϕ using the last samples of Z according to

$$\theta_{dk} = \frac{N_{dk} + \alpha}{N_d + \alpha \cdot K} \tag{A.14}$$

$$\phi_{kv} = \frac{N_{kv} + \beta_{kv_k}^*}{\sum_v N_{kv} + \beta_{kv}^*} \tag{A.15}$$

In the empirical study, we repeat the above MCMC process 50,000 times and then use 25,000 samples (excluding burn-in samples) to calculate the posterior means and intervals of HPD. The settings of prior distribution used in the empirical study are as follows:

$$\begin{aligned}
\theta_d &\sim Dirichlet(\alpha), \quad \alpha_k = 1.0\ \forall k\\
\phi_k &\sim Dirichlet(\Lambda^{(k)}\beta), \quad \beta_v = 1.0\ \forall v\\
\gamma_s &\sim N(0, g_{s,0}^{-1}\cdot I), \quad g_{s,0} = 0.1\\
\delta_s &\sim N(0, d_{s,0}^{-1}\cdot I), \quad d_{s,0} = 0.1\\
\gamma_h &\sim N(0, g_{h,0}^{-1}\cdot I), \quad g_{h,0} = 0.1\\
\delta_h &\sim N(0, d_{h,0}^{-1}\cdot I), \quad d_{h,0} = 0.1
\end{aligned}$$

A.4 Estimation Procedure of the Supervised-Sentiment LDA2vec model with Brand Heterogeneity and Consumer Attributes

In this section, we describe the update procedure for word and topic embedding vectors using the gradient-based stochastic optimization, and then the estimation procedure for the topic assignments and regression coefficients using the MCMC sampler given the optimized embedding vectors.

First, we introduce the state-of-the-art stochastic optimization method for neural network based models, which is called Adam (Kingma and Ba, 2015). From the model likelihood of the word embedding part (5.2), given the topic assignment for each word, the loss function L can be defined as follows.

$$L = \sum_{ij} L_{ij}, \quad L_{ij} = \log \sigma(\overrightarrow{c}_i^{\top}\overrightarrow{w}_j) + \sum_{n\sim P_n(w)}^{N} \sigma(-\overrightarrow{c}_i^{\top}\overrightarrow{w}_n). \tag{A.16}$$

Let $g_i^{(t)}$ be the gradient function for embedding vector of word i at time t, it is defined as $g_i^{(t)} = \eta \times \frac{\partial L}{\partial w_i}(\overrightarrow{w}_i^{(t)})$, where η is the learning rate parameter and set to $\eta = 0.001$ in this study. To obtain the updated vector $\overrightarrow{w}_i^{(t+1)}$ from the previous vector $\overrightarrow{w}_i^{(t+1)}$, we use the following Adam algorithm.

$$\begin{aligned}
m_i^{(t)} &= \beta_1 \times m_i^{(t-1)} + (1-\beta_1)\times g_i^{(t)}, \qquad m_i^{(0)} = 0.0\\
v_i^{(t)} &= \beta_2 \times v_i^{(t-1)} + (1-\beta_2)\times \left(g_i^{(t)}\right)^2, \qquad v_i^{(0)} = 0.0\\
\overrightarrow{w}_i^{(t+1)} &= \overrightarrow{w}_i^{(t)} - \frac{\eta}{\sqrt{\hat{v}_i^{(t)}} + \epsilon} \times \hat{m}_i^{(t)},
\end{aligned} \tag{A.17}$$

where β_1 and β_2 are decay parameters and set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, and ϵ is small value for avoiding zero-dividing and set to $\epsilon = 10^{-8}$ according to the original paper's suggestions, respectively. Also, $\hat{m}_i^{(t)}$ and $\hat{v}_i^{(t)}$ are corrections of the bias from the mean and defined as $m_i^{(t)} = \frac{m_i^{(t)}}{1-\beta_1^t}$ and $\hat{v}_i^{(t)} = \frac{v_i^{(t)}}{1-\beta_2^t}$, respectively. Topic vectors $\overrightarrow{t}_k$ are also updated in the similar way.

In this study, we use the free-available[1] word embedding representations (GloVe, Pennington, Socher, and Manning, 2014) pre-trained by Wikipedia large corpus as the initial values of the word embedding vectors. In general, while the pre-trained models contribute to improve the predictive performance and the faster optimization, this study expects it to contribute to improve the topic interpretability. It is because the topic assignments by considering the general meanings of the words will allow us to appropriately and quickly optimize the model such that it encompasses the unique and non-general meanings in the focal domain, compared to initialization in the absence of any prior information.

For the preference measurement part of the proposed model, we omit deriving the posterior distributions in detail because it has a similar structure to the ordered probit model of Chapter 4, except for the considerations of the brand heterogeneity and the effects of consumer attributes. Therefore, the algorithm of the hybrid approach for estimating the proposed model, with the stochastic optimization for embedding vectors and the MCMC sampling for the preference measurement model parameters, is as follows:

1. initialize embedding vectors using pre-trained vectors and preference measurement model parameters randomly
2. iterate the following optimizer and sampler for the pre-determined times
   (a) update $\overrightarrow{w}_i$ and $\overrightarrow{t}_k$ using the Adam algorithm (A.17)
   (b) update topic distribution (θ) in the Metropolis-Hasting fashion
   (c) sample topic assignment (Z), thresholds (τ), latent continuous rating score (y), and regression parameters (α, β, γ, and σ)

[1] It can be downloaded from the Stanford University's NLP project website `https://nlp.stanford.edu/projects/glove/`.

where we stop the optimization of the embedding vectors before other parameters not converged for avoiding the overfit to the training dataset, and this optimization technique is called early-stopping and commonly used in the machine learning field.

In the empirical study, we conduct the above estimation algorithm with 1,000 times MCMC sampler (early stop the embedding vector optimization at 100 times), and then last 800 samples to calculate the posterior means and intervals of HPD. The settings of prior distribution used in the empirical study are as follows:

$$\begin{aligned}
\theta_d &\sim Dirichlet(\theta_0), \quad \theta_0 = 0.8 \\
\alpha_b &\sim N(0, \sigma^2_{\alpha 0}), \quad \sigma^2_{\alpha 0} = 100.0 \\
\beta_{bk} &\sim N(0, \sigma^2_{\beta 0}), \quad \sigma^2_{\beta 0} = 100.0 \\
\gamma_q &\sim N(0, \sigma^2_{\gamma 0}), \quad \sigma^2_{\gamma 0} = 100.0 \\
\sigma^2 &\sim inverse - Gamma(n_0, s_0), \quad n_0 = s_0 = 1.0
\end{aligned}$$

Appendix B

Definitions of Information Criterion

B.1 Definition of WAIC for the MMSTB

The definition of WAIC for the MMSTB is as follows:

$$\begin{aligned} lpd^{(i)} &= \log\left(\frac{1}{G}\sum_{g=b+1}^{G}\prod_{j=1}^{D}P\left(a_{ij}|H^{(g)},\Psi^{(g)}\right)\right. \\ &\qquad \left.\prod_{m=1}^{M_i}P\left(w_{im}|H^{(g)},\Theta^{(g)},\Phi^{(g)}\right)\right) \qquad \text{(B.1)} \\ P_{waic}^{(i)} &= \frac{G}{G-1}\left(\frac{1}{G}\sum_{g=b+1}^{G}\left(\sum_{j=1}^{D}\log P\left(a_{ij}|H^{(g)},\Psi^{(g)}\right)^2\right.\right. \\ &\qquad \left.+\sum_{m=1}^{M_i}\log P\left(w_{im}|H^{(g)},\Theta^{(g)},\Phi^{(g)}\right)^2\right) \\ &\qquad -\left(\frac{1}{G}\sum_{g=b+1}^{G}\left(\sum_{j=1}^{D}\log P\left(a_{ij}|H^{(g)},\Psi^{(g)}\right)\right.\right. \\ &\qquad \left.\left.\left.+\sum_{m=1}^{M_i}\log P\left(w_{im}|H^{(g)},\Theta^{(g)},\Phi^{(g)}\right)\right)\right)^2\right) \qquad \text{(B.2)} \end{aligned}$$

$$WAIC = -2\sum_{i=1}^{D}\left(lpd^{(i)} - P_{waic}^{(i)}\right), \qquad \text{(B.3)}$$

where $P\left(a_{ij}|H^{(g)},\Psi^{(g)}\right)$ and $P\left(w_{im}|H^{(g)},\Theta^{(g)},\Phi^{(g)}\right)$ are the model likelihood conditioned with the parameters estimated using samples at sth iteration

$$P\left(a_{ij}|H^{(g)},\Psi^{(g)}\right) = \sum_{k=1}^{K}\sum_{k'=1}^{K}\eta_{ik}\cdot\eta_{jk'}^{(g)}\cdot\psi_{kk'}^{(g)\mathbb{I}(a_{ij}=1)}\cdot(1-\psi_{kk'})^{(g)\mathbb{I}(a_{ij}=0)} \tag{B.4}$$

$$P\left(w_{im}|H^{(g)},\Theta^{(g)},\Phi^{(g)}\right) = \sum_{k=1}^{K}\sum_{l=1}^{L}\eta_{ik}^{(g)}\cdot\theta_{kl}^{(g)}\cdot\phi_{lw_{im}}^{(g)}. \tag{B.5}$$

Bibliography

Airoldi, Edoardo M. et al. (2008). “Mixed membership stochastic blockmodels”. In: *Journal of Machine Learning Research* 9.SEP, pp. 1981–2014.

Ameri, Mina, Elisabeth Honka, and Ying Xie (2019). “Word of Mouth, Observed Adoptions, and Anime-Watching Decisions: The Role of the Personal vs. the Community Network”. In: *Marketing Science* 38.4, pp. 567–583. ISSN: 0732-2399. DOI: 10.1287/mksc.2019.1155. URL: http://pubsonline.informs.org/doi/10.1287/mksc.2019.1155.

Andersen, Barbara Vad and Grethe Hyldig (2015). “Food satisfaction: Integrating feelings before, during and after food intake”. In: *Food Quality and Preference* 43, pp. 126–134. ISSN: 0950-3293. DOI: 10.1016/GOODQUAL.2015.03.004.

Ansari, Asim et al. (2018). “Building a social network for success”. In: *Journal of Marketing Research* 55.3, pp. 321–338. ISSN: 00222437. DOI: 10.1509/jmr.12.0417.

Aral, Sinan, Lev Muchnik, and Arun Sundararajan (2009). “Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks”. In: *Proceedings of the National Academy of Sciences* 106.51, pp. 21544–21549. ISSN: 0027-8424. DOI: 10.1073/pnas.0908800106. URL: https://www-pnas-org.proxy.lib.uiowa.edu/content/pnas/106/51/21544.full.pdfhttp://www.pnas.org/cgi/doi/10.1073/pnas.0908800106.

Archak, Nikolay, Anindya Ghose, and Panagiotis G. Ipeirotis (2011). “Deriving the Pricing Power of Product Features by Mining Consumer Reviews”. In: *Management Science* 57.8, pp. 1485–1509. ISSN: 0025-1909. DOI: 10.1287/mnsc.1110.1370. URL: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1110.1370.

Barbillon, Pierre et al. (2017). “Stochastic block models for multiplex networks: an application to a multilevel network of researchers”. In: *Journal of the Royal Statistical Society. Series A: Statistics in Society* 180.1, pp. 295–314. ISSN: 1467985X. DOI: 10.1111/rssa.12193. arXiv: 1501.06444.

Bass, Frank M. (1969). "A New Product Growth for Model Consumer Durables". In: *Management Science* 15.5, pp. 215–227. URL: http://www.jstor.org/stable/2628128.

Berger, Jonah et al. (2020). "Uniting the Tribes: Using Text for Marketing Insight". In: *Journal of Marketing* 84.1, pp. 1–25. ISSN: 15477185. DOI: 10.1177/0022242919873106.

Bhattacharya, Anirban et al. (2015). "Dirichlet–Laplace Priors for Optimal Shrinkage". In: *Journal of the American Statistical Association* 110.512, pp. 1479–1490. ISSN: 1537274X. DOI: 10.1080/01621459.2014.960967. arXiv: 1401.5398.

Bhattacharya, Prasanta et al. (2019). "A Coevolution Model of Network Structure and User Behavior: The Case of Content Generation in Online Social Networks". In: *Information Systems Research* 30.1, pp. 117–132. ISSN: 1047-7047. DOI: 10.1287/isre.2018.0790. URL: http://pubsonline.informs.org/doi/10.1287/isre.2018.0790.

Bi, Jian-Wu et al. (2019). "Wisdom of crowds: Conducting importance-performance analysis (IPA) through online reviews". In: *Tourism Management* 70, pp. 460–478. ISSN: 02615177. DOI: 10.1016/j.tourman.2018.09.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S0261517718302188.

Blei, David M. and John D. Lafferty (2005). "Correlated topic models". In: *Advances in Neural Information Processing Systems*, pp. 147–154. ISSN: 10495258.

— (2006). "Dynamic topic models". In: *ACM International Conference Proceeding Series* 148, pp. 113–120. DOI: 10.1145/1143844.1143859.

Blei, David M. and Jon D. McAuliffe (2007). "Supervised topic models". In: *Advances in neural information processing systems*, pp. 121–128.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet allocation". In: *Journal of Machine Learning Research* 3.4-5, pp. 993–1022. ISSN: 15324435.

Bollinger, Bryan and Kenneth Gillingham (2012). "Peer effects in the diffusion of solar photovoltaic panels". In: *Marketing Science* 31.6, pp. 900–912. ISSN: 07322399. DOI: 10.1287/mksc.1120.0727.

Bouveyron, C., P. Latouche, and R. Zreik (2018). "The stochastic topic block model for the clustering of vertices in networks with textual edges". In: *Statistics and Computing* 28.1, pp. 11–31. ISSN: 15731375. DOI: 10.1007/s11222-016-9713-7.

Burt, Roland D. (2001). *Bandwidth and echo: Trust, information, and gossip in social networks.*

Büschken, Joachim and Greg M. Allenby (2016). "Sentence-based text analysis for customer reviews". In: *Marketing Science* 35.6, pp. 953–975. ISSN: 1526548X. DOI: `10.1287/mksc.2016.0993`.

— (2020). "Improving Text Analysis Using Sentence Conjunctions and Punctuation". In: *Marketing Science* 39.4, pp. 727–742. ISSN: 0732-2399. DOI: `10.1287/mksc.2019.1214`. URL: `http://pubsonline.informs.org/doi/10.1287/mksc.2019.1214`.

Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott (2009). "Handling Sparsity via the Horseshoe". In: *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, pp. 73–80.

Caselles-Dupré, Hugo, Florian Lesaint, and Jimena Royo-Letelier (2018). "Word2vec applied to recommendation". In: *Proceedings of the 12th ACM Conference on Recommender Systems*. New York, NY, USA: ACM, pp. 352–356. ISBN: 9781450359016. DOI: `10.1145/3240323.3240377`. URL: `https://dl.acm.org/doi/10.1145/3240323.3240377`.

Cater, C. K. and R. Kohn (1994). "On Gibbs sampling for state space models". In: *Biometrika* 81.3, pp. 541–553. ISSN: 0006-3444. DOI: `10.1093/biomet/81.3.541`. URL: `https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/81.3.541`.

Chang, Jonathan and David M. Blei (2010). "Hierarchical relational models for document networks". In: *The Annals of Applied Statistics* 4.1, pp. 124–150. ISSN: 1932-6157. DOI: `10.1214/09-AOAS309`. URL: `http://projecteuclid.org/euclid.aoas/1273584450`.

Chang, Jonathan et al. (2009). "Reading tea leaves: How humans interpret topic models". In: *Advances in Neural Information Processing Systems*, pp. 288–296.

Chen, Kehui and Jing Lei (2018). "Network Cross-Validation for Determining the Number of Communities in Network Data". In: *Journal of the American Statistical Association* 113.521, pp. 241–251. ISSN: 1537274X. DOI: `10.1080/01621459.2016.1246365`. URL: `https://doi.org/10.1080/01621459.2016.1246365`.

Chen, Xi, Ralf Van der Lans, and Tuan Q. Phan (2017). "Uncovering the Importance of Relationship Characteristics in Social Networks: Implications for Seeding Strategies". In: *Journal of Marketing Research* 54.2, pp. 187–201. ISSN: 0022-2437. DOI: 10.1509/jmr.12.0511. URL: http://journals.sagepub.com/doi/10.1509/jmr.12.0511.

Chen, Yubo, Qi Wang, and Jinhong Xie (2011). "Online Social Interactions: A Natural Experiment on Word of Mouth versus Observational Learning". In: *Journal of Marketing Research* 48.2, pp. 238–254. ISSN: 0022-2437. DOI: 10.1509/jmkr.48.2.238. URL: http://journals.sagepub.com/doi/10.1509/jmkr.48.2.238.

Chen, Yubo and Jinhong Xie (2008). "Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix". In: *Management Science* 54.3, pp. 477–491. ISSN: 0025-1909. DOI: 10.1287/mnsc.1070.0810. URL: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1070.0810.

Chevalier, Judith A. and Dina Mayzlin (2006). "The effect of word of mouth on sales: Online book reviews". In: *Journal of Marketing Research* 43.3, pp. 345–354. ISSN: 00222437. DOI: 10.1509/jmkr.43.3.345.

Chi, Christina Geng-Qing et al. (2013). "Investigating the Structural Relationships Between Food Image, Food Satisfaction, Culinary Quality, and Behavioral Intentions: The Case of Malaysia". In: *International Journal of Hospitality & Tourism Administration* 14.2, pp. 99–120. ISSN: 1525-6480. DOI: 10.1080/15256480.2013.782215. URL: http://www.tandfonline.com/doi/abs/10.1080/15256480.2013.782215.

Chintagunta, Pradeep K., Shyam Gopinath, and Sriram Venkataraman (2010). "The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets". In: *Marketing Science* 29.5, pp. 944–957. ISSN: 07322399. DOI: 10.1287/mksc.1100.0572.

Cho, Youngsang, Junseok Wang, and Daeho Lee (2012). "Identification of effective opinion leaders in the diffusion of technological innovation: A social network approach". In: *Technological Forecasting and Social Change* 79.1, pp. 97–106. ISSN: 00401625. DOI: 10.1016/j.techfore.2011.06.003. URL: http://dx.doi.org/10.1016/j.techfore.2011.06.003.

Choi, Hanool, Sang Hoon Kim, and Jeho Lee (2010). "Role of network structure and network effects in diffusion of innovations". In: *Industrial Marketing Management* 39.1, pp. 170–177. ISSN: 00198501. DOI: 10.1016/j.indmarman.2008.08.006. URL: http://dx.doi.org/10.1016/j.indmarman.2008.08.006.

Daudin, J.-J., F. Picard, and S. Robin (2008). "A mixture model for random graphs". In: *Statistics and Computing* 18.2, pp. 173–183. ISSN: 0960-3174. DOI: 10.1007/s11222-007-9046-7. URL: http://link.springer.com/10.1007/s11222-007-9046-7.

Decker, Reinhold and Michael Trusov (2010). "Estimating aggregate consumer preferences from online product reviews". In: *International Journal of Research in Marketing* 27.4, pp. 293–307. ISSN: 01678116. DOI: 10.1016/j.ijresmar.2010.09.001. URL: http://dx.doi.org/10.1016/j.ijresmar.2010.09.001.

Delre, Sebastiano A., Wander Jager, and Marco A. Janssen (2007). "Diffusion dynamics in small-world networks with heterogeneous consumers". In: *Computational and Mathematical Organization Theory* 13.2, pp. 185–202. ISSN: 1381-298X. DOI: 10.1007/s10588-006-9007-2. URL: http://link.springer.com/10.1007/s10588-006-9007-2.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Association for Computational Linguistics*, pp. 4171–4186.

Dover, Yaniv, Jacob Goldenberg, and Daniel Shapira (2012). "Network traces on penetration: Uncovering degree distribution from adoption data". In: *Marketing Science* 31.4, pp. 689–712. ISSN: 07322399. DOI: 10.1287/mksc.1120.0711.

Felbermayr, Armin and Alexandros Nanopoulos (2016). "The Role of Emotions for the Perceived Usefulness in Online Customer Reviews". In: *Journal of Interactive Marketing* 36, pp. 60–76. ISSN: 15206653. DOI: 10.1016/j.intmar.2016.05.004. URL: http://dx.doi.org/10.1016/j.intmar.2016.05.004.

Fischer, Gregory W. et al. (1999). "Goal-Based Construction of Preferences: Task Goals and the Prominence Effect". In: *Management Science* 45.8, pp. 1057–1075. ISSN: 0025-1909. DOI: 10.1287/mnsc.45.8.1057. URL: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.45.8.1057.

Franses, Philip Hans (2005). "On the Use of Econometric Models for Policy Simulation in Marketing". In: *Journal of Marketing Research* 42.1, pp. 4–14. ISSN: 0022-2437. DOI: 10.1509/jmkr.42.1.4.56891. URL: http://journals.sagepub.com/doi/10.1509/jmkr.42.1.4.56891.

Gelman, Andrew et al. (2013). *Bayesian Data Analysis, Third Edition*. New York: Chapman and Hall/CRC, p. 675. ISBN: 9781439898208. DOI: 10.1201/b16018. URL: https://www.taylorfrancis.com/books/9781439898208.

Ghose, Anindya, Panagiotis G. Ipeirotis, and Beibei Li (2012). "Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content". In: *Marketing Science* 31.3, pp. 493–520. ISSN: 0732-2399. DOI: 10.1287/mksc.1110.0700. URL: http://pubsonline.informs.org/doi/abs/10.1287/mksc.1110.0700.

Ghose, Anindya, Panagiotis G. Ipeirotis, and Arun Sundararajan (2007). "Opinion Mining Using Econometrics: A Case Study on Reputation Systems". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 416–423.

Gilbert, C. H. E. and Erric Hutto (2014). "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *International Conference on Weblogs and Social Media*.

Glynn, Chris et al. (2019). "Bayesian Analysis of Dynamic Linear Topic Models". In: *Bayesian Analysis* 14.1, pp. 53–80. ISSN: 1936-0975. DOI: 10.1214/18-ba1100.

Godes, David and Dina Mayzlin (2004). "Using online conversations to study word-of-mouth communication". In: *Marketing Science* 23.4. ISSN: 07322399. DOI: 10.1287/mksc.1040.0071.

Gong, Shiyang et al. (2017). "Tweeting as a marketing tool: A field experiment in the TV industry". In: *Journal of Marketing Research* 54.6, pp. 833–850. ISSN: 00222437. DOI: 10.1509/jmr.14.0348.

Granovetter, Mark S. (1973). "The Strength of Weak Ties". In: *American Journal of Sociology* 78.6, pp. 1360–1380. ISSN: 0002-9602. DOI: 10.1086/225469. URL: https://www.journals.uchicago.edu/doi/10.1086/225469.

Greene, Derek and Pádraig Cunningham (2006). "Practical solutions to the problem of diagonal dominance in kernel document clustering". In: *Proceedings of the 23rd* *international conference on Machine learning - ICML '06*. New York, New York, USA: ACM Press, pp. 377–384. ISBN: 1595933832. DOI: `10.1145/1143844.1143892`. URL: `http://portal.acm.org/citation.cfm?doid=1143844.1143892`.

Griffiths, T. L. and M. Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.Supplement 1, pp. 5228–5235. ISSN: 0027-8424. DOI: `10.1073/pnas.0307752101`. URL: `http://www.pnas.org/cgi/doi/10.1073/pnas.0307752101`.

Haenlein, Michael (2013). "Social interactions in customer churn decisions: The impact of relationship directionality". In: *International Journal of Research in Marketing* 30.3, pp. 236–248. ISSN: 01678116. DOI: `10.1016/j.ijresmar.2013.03.003`. URL: `http://dx.doi.org/10.1016/j.ijresmar.2013.03.003`.

Handcock, Mark S., Adrian E. Raftery, and Jeremy M. Tantrum (2007). "Model-based clustering for social networks". In: *Journal of the Royal Statistical Society. Series A: Statistics in Society* 170.2, pp. 301–354. ISSN: 09641998. DOI: `10.1111/j.1467-985X.2007.00471.x`.

Hartmann, Wesley R. (2010). "Demand estimation with social interactions and the implications for targeted marketing". In: *Marketing Science* 29.4, pp. 585–601. ISSN: 07322399. DOI: `10.1287/mksc.1100.0559`.

Hartmann, Wesley R. et al. (2008). "Modeling social interactions: Identification, empirical methods and policy implications". In: *Marketing Letters* 19.3-4, pp. 287–304. ISSN: 09230645. DOI: `10.1007/s11002-008-9048-z`.

Ho-Dac, Nga N., Stephen J. Carson, and William L. Moore (2013). "The Effects of Positive and Negative Online Customer Reviews: Do Brand Strength and Category Maturity Matter?" In: *Journal of Marketing* 77.6, pp. 37–53. ISSN: 0022-2429. DOI: `10.1509/jm.11.0011`. URL: `http://journals.sagepub.com/doi/10.1509/jm.11.0011`.

Hoeffler, Steve (2003). "Measuring Preferences for Really New Products". In: *Journal of Marketing Research* 40.4, pp. 406–420. ISSN: 0022-2437. DOI: `10.1509/jmkr.40.4.406.19394`. URL: `http://journals.sagepub.com/doi/10.1509/jmkr.40.4.406.19394`.

Hoff, Peter D., Adrian E. Raftery, and Mark S. Handcock (2002). "Latent space approaches to social network analysis". In: *Journal of the American Statistical Association* 97.460, pp. 1090–1098. ISSN: 01621459. DOI: 10.1198/016214502388618906.

Hoffman, Matthew, Francis R. Bach, and David M. Blei (2010). "Online Learning for Latent Dirichlet Allocation". In: *Advances in Neural Information Processing Systems*.

Hoffman, Thomas (1999). "Probabilistic latent semantic indexing". In: *International ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57.

Hou, Tianjun et al. (2019). "Mining Changes in User Expectation Over Time From Online Reviews". In: *Journal of Mechanical Design* 141.9. ISSN: 1050-0472. DOI: 10.1115/1.4042793. URL: https://asmedigitalcollection.asme.org/mechanicaldesign/article/doi/10.1115/1.4042793/727242/Mining-Changes-in-User-Expectation-Over-Time-From.

Hubert, Lawrence and Phipps Arabie (1985). "Comparing partitions". In: *Journal of Classification* 2.1, pp. 193–218. ISSN: 0176-4268. DOI: 10.1007/BF01908075. URL: http://link.springer.com/10.1007/BF01908075.

Igarashi, Mirai and Nobuhiko Terui (2020). "Characterization of topic-based online communities by combining network data and user generated content". In: *Statistics and Computing*. ISSN: 0960-3174. DOI: 10.1007/s11222-020-09947-5. URL: http://link.springer.com/10.1007/s11222-020-09947-5.

Iyengar, Raghuram, Christophe Van den Bulte, and Thomas W. Valente (2011). "Opinion leadership and social contagion in new product diffusion". In: *Marketing Science* 30.2, pp. 195–212. ISSN: 07322399. DOI: 10.1287/mksc.1100.0566.

Jeong, H. et al. (2001). "Lethality and centrality in protein networks". In: *Nature* 411.6833, pp. 41–42. ISSN: 0028-0836. DOI: 10.1038/35075138. URL: http://www.nature.com/articles/35075138.

Kano, Noriaki et al. (1984). "Attractive Quality and Must-Be Quality". In: *Journal of The Japanese Society for Quality Control* 14.2, pp. 147–156. DOI: 10.20684/quality.14.2_147.

Karrer, Brian and M. E.J. Newman (2011). "Stochastic blockmodels and community structure in networks". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 83.1, pp. 1–10. ISSN: 15393755. DOI: 10.1103/PhysRevE.83.016107.

Keeling, Matt (2005). "The implications of network structure for epidemic dynamics". In: *Theoretical Population Biology* 67.1, pp. 1–8. ISSN: 00405809. DOI: 10.1016/j.tpb.2004.08.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0040580904001121.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *Proceedings of the 3rd International Conference for Learning Representations*.

Krebs, Validis E. (2002). "Mapping Networks of Terrorist Cells". In: *Connections* 24.3, pp. 43–52.

Krivitsky, Pavel N. et al. (2009). "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models". In: *Social Networks* 31.3, pp. 204–213. ISSN: 03788733. DOI: 10.1016/j.socnet.2009.04.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378873309000173.

Küster, Inés and Natalia Vila (2017). "Health/Nutrition food claims and low-fat food purchase: Projected personality influence in young consumers". In: *Journal of Functional Foods* 38, pp. 66–76. ISSN: 17564646. DOI: 10.1016/j.jff.2017.08.046. URL: https://linkinghub.elsevier.com/retrieve/pii/S175646461730511X.

Lanz, Andreas et al. (2019). "Climb or Jump: Status-Based Seeding in User-Generated Content Networks". In: *Journal of Marketing Research* 56.3, pp. 361–378. ISSN: 0022-2437. DOI: 10.1177/0022243718824081. URL: http://journals.sagepub.com/doi/10.1177/0022243718824081.

Latouche, P, E Birmelé, and C Ambroise (2012). "Variational Bayesian inference and complexity control for stochastic block models". In: *Statistical Modelling: An International Journal* 12.1, pp. 93–115. ISSN: 1471-082X. DOI: 10.1177/1471082X1001200105. URL: http://journals.sagepub.com/doi/10.1177/1471082X1001200105.

Latouche, Pierre, Etienne Birmelé, and Christophe Ambroise (2011). "Overlapping stochastic block models with application to the French political blogosphere". In: *The Annals of Applied Statistics* 5.1, pp. 309–336. ISSN: 1932-6157. DOI: 10.1214/10-AOAS382. URL: http://projecteuclid.org/euclid.aoas/1300715192.

Lee, Thomas Y. and Eric T. Bradlow (2011). "Automated marketing research using online customer reviews". In: *Journal of Marketing Research* 48.5, pp. 881–894. ISSN: 00222437. DOI: 10.1509/jmkr.48.5.881.

Leskovec, Jure, Lada A. Adamic, and Bernardo A. Huberman (2007). "The dynamics of viral marketing". In: *ACM Transactions on the Web* 1.1, p. 5. ISSN: 1559-1131. DOI: 10.1145/1232722.1232727. URL: https://dl.acm.org/doi/10.1145/1232722.1232727.

Libai, Barak, Eitan Muller, and Renana Peres (2013). "Decomposing the Value of Word-of-Mouth Seeding Programs: Acceleration Versus Expansion". In: *Journal of Marketing Research* 50.2, pp. 161–176.

Lin, Chenghua et al. (2012). "Weakly supervised joint sentiment-topic detection from text". In: *IEEE Transactions on Knowledge and Data Engineering* 24.6, pp. 1134–1145. ISSN: 10414347. DOI: 10.1109/TKDE.2011.48.

Liu, Xiaoming et al. (2005). "Co-authorship networks in the digital library research community". In: *Information Processing & Management* 41.6, pp. 1462–1480. ISSN: 03064573. DOI: 10.1016/j.ipm.2005.03.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306457305000336.

Liu, Yan, Alexandru Niculescu-Mizil, and Wojciech Gryc (2009). "Topic-link LDA: Joint models of topic and author community". In: *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, pp. 665–672.

Liu, Yong (2006). "Word of mouth for movies: Its dynamics and impact on box office revenue". In: *Journal of Marketing* 70.3, pp. 74–89. ISSN: 00222429. DOI: 10.1509/jmkg.70.3.74.

Lu, Shuya, Jianan Wu, and Shih Lun (Allen) Tseng (2018). "How Online Reviews Become Helpful: A Dynamic Perspective". In: *Journal of Interactive Marketing* 44, pp. 17–28. ISSN: 15206653. DOI: 10.1016/j.intmar.2018.05.005. URL: https://doi.org/10.1016/j.intmar.2018.05.005.

Ludwig, Stephan et al. (2013). "More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates". In: *Journal of Marketing* 77.1, pp. 87–103. ISSN: 15477185. DOI: 10.1509/jm.11.0560.

Manski, Charles F. (1993). "Identification of Endogenous Social Effects: The Reflection Problem". In: *The Review of Economic Studies* 60.3, p. 531. ISSN: 00346527. DOI:

10.2307/2298123. URL: http://www.cmap.polytechnique.fr/{~}rama/ehess/manski3.pdfhttps://academic.oup.com/restud/article-lookup/doi/10.2307/2298123.

Mason, Michela Cesarina and Federico Nassivera (2013). "A Conceptualization of the Relationships Between Quality, Satisfaction, Behavioral Intention, and Awareness of a Festival". In: *Journal of Hospitality Marketing & Management* 22.2, pp. 162–182. ISSN: 1936-8623. DOI: 10.1080/19368623.2011.643449. URL: http://www.tandfonline.com/doi/abs/10.1080/19368623.2011.643449.

Matias, Catherine and Vincent Miele (2017). "Statistical clustering of temporal networks through a dynamic stochastic block model". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 79.4, pp. 1119–1141. ISSN: 14679868. DOI: 10.1111/rssb.12200.

Mauri, Aurelio G. and Roberta Minazzi (2013). "Web reviews influence on expectations and purchasing intentions of hotel potential customers". In: *International Journal of Hospitality Management* 34, pp. 99–107. ISSN: 02784319. DOI: 10.1016/j.ijhm.2013.02.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S0278431913000236.

McDaid, Aaron F. et al. (2013). "Improved Bayesian inference for the stochastic block model with application to large networks". In: *Computational Statistics & Data Analysis* 60, pp. 12–31. ISSN: 01679473. DOI: 10.1016/j.csda.2012.10.021. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167947312003891.

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems*, pp. 3111–3119. ISBN: 9781945626258. DOI: 10.18653/v1/d16-1146. arXiv: 1606.08359.

Mimno, David et al. (2011). "Optimizing semantic coherence in topic models". In: *Empirical Methods in Natural Language Processing* 2, pp. 262–272.

Moe, Wendy W., Michael Trusov, and Robert H. Smith (2011). "The value of social dynamics in online product ratings forums". In: *Journal of Marketing Research* 48.3, pp. 444–456. ISSN: 00222437. DOI: 10.1509/jmkr.48.3.444.

Moody, Christopher E (2016). “Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec”. In: arXiv: 1605.02019. URL: http://arxiv.org/abs/1605.02019.

Moon, Sangkil and Wagner A. Kamakura (2017). “A picture is worth a thousand words: Translating product reviews into a product positioning map”. In: *International Journal of Research in Marketing* 34.1, pp. 265–285. ISSN: 01678116. DOI: 10.1016/j.ijresmar.2016.05.007. URL: http://dx.doi.org/10.1016/j.ijresmar.2016.05.007.

Mudambi, Susan M. and David Schuff (2010). “What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com”. In: *MIS Quarterly: Management Information Systems* 34.1, pp. 185–200.

Mukherjee, Prithwiraj (2014). “How chilling are network externalities? The role of network structure”. In: *International Journal of Research in Marketing* 31.4, pp. 452–456. ISSN: 01678116. DOI: 10.1016/j.ijresmar.2014.09.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167811614000779.

Muller, Eitan and Renana Peres (2019). “The effect of social networks structure on innovation performance: A review and directions for research”. In: *International Journal of Research in Marketing* 36.1, pp. 3–19. ISSN: 01678116. DOI: 10.1016/j.ijresmar.2018.05.003. URL: https://doi.org/10.1016/j.ijresmar.2018.05.003.

Nair, Harikesh S., Puneet Manchanda, and Tulikaa Bhatia (2010). “Asymmetric Social Interactions in Physician Prescription Behavior: The Role of Opinion Leaders”. In: *Journal of Marketing Research* 47.5, pp. 883–895. ISSN: 0022-2437. DOI: 10.1509/jmkr.47.5.883. URL: http://journals.sagepub.com/doi/10.1509/jmkr.47.5.883.

Nam, Sungjoon, Puneet Manchanda, and Pradeep K. Pradeep (2010). “The effect of signal quality and contiguous word of mouth on customer acquisition for a video-on-demand service”. In: *Marketing Science* 29.4, pp. 690–700. ISSN: 07322399. DOI: 10.1287/mksc.1090.0550.

Narayan, Vishal, Vithala R. Rao, and Carolyne Saunders (2011). “How peer influence affects attribute preferences: A Bayesian updating mechanism”. In: *Marketing Science* 30.2, pp. 368–384. ISSN: 07322399. DOI: 10.1287/mksc.1100.0618.

Netzer, Oded et al. (2008). "Beyond conjoint analysis: Advances in preference measurement". In: *Marketing Letters* 19.3-4, pp. 337–354. ISSN: 09230645. DOI: 10.1007/s11002-008-9046-1.

Netzer, Oded et al. (2012). "Mine Your Own Business: Market-Structure Surveillance Through Text Mining". In: *Marketing Science* 31.3, pp. 521–543. ISSN: 0732-2399. DOI: 10.1287/mksc.1120.0713. URL: http://pubsonline.informs.org/doi/abs/10.1287/mksc.1120.0713.

Newman, M. E. J. (2006). "Modularity and community structure in networks". In: *Proceedings of the National Academy of Sciences* 103.23, pp. 8577–8582. ISSN: 0027-8424. DOI: 10.1073/pnas.0601602103. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0601602103.

Newton, Michael A. and Adrian E. Raftery (1994). "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 56.1, pp. 3–26. ISSN: 00359246. DOI: 10.1111/j.2517-6161.1994.tb01956.x. URL: http://doi.wiley.com/10.1111/j.2517-6161.1994.tb01956.x.

Ng, Andrew Y., Michael I. Jordan, and Yair Weiss (2002). "On spectral clustering: Analysis and an algorithm". In: *Advances in Neural Information Processing Systems*, pp. 849–856.

Nowicki, Krzysztof and Tom A.B. Snijders (2001). "Estimation and prediction for stochastic blockstructures". In: *Journal of the American Statistical Association* 96.455, pp. 1077–1087. ISSN: 1537274X. DOI: 10.1198/016214501753208735.

Park, Eunho et al. (2018). "Social dollars in online communities: The effect of product, user, and network characteristics". In: *Journal of Marketing* 82.1, pp. 93–114. ISSN: 15477185. DOI: 10.1509/jm.16.0271.

Park, Trevor and George Casella (2008). "The Bayesian Lasso". In: *Journal of the American Statistical Association* 103.482, pp. 681–686. ISSN: 0162-1459. DOI: 10.1198/016214508000000337. URL: https://www.tandfonline.com/doi/full/10.1198/016214508000000337.

Pathak, Nishth et al. (2008). "Social Topic Models for Community Extraction". In: *SNA-KDD workshop*, p. 2008.

Peng, Jing et al. (2018). "Network overlap and content sharing on social media platforms". In: *Journal of Marketing Research* 55.4, pp. 571–585. ISSN: 00222437. DOI: 10.1509/jmr.14.0643.

Pennington, Jeffery, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation". In: *Empirical Methods in Natural Language Processing2*, pp. 1532–1543.

Peres, Renana (2014). "The impact of network characteristics on the diffusion of innovations". In: *Physica A: Statistical Mechanics and its Applications* 402, pp. 330–343. ISSN: 03784371. DOI: 10.1016/j.physa.2014.02.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378437114000946.

Peres, Renana and Christophe Van den Bulte (2014). "When to take or forgo new product exclusivity: Balancing protection from competition against word-of-mouth spillover". In: *Journal of Marketing* 78.2, pp. 83–100. ISSN: 15477185. DOI: 10.1509/jm.12.0344.

Phan, Tuan Q. and David Godes (2018). "The evolution of influence through endogenous link formation". In: *Marketing Science* 37.2, pp. 259–278. ISSN: 1526548X. DOI: 10.1287/mksc.2017.1077.

Polson, Nicholas G., James G. Scott, and Jesse Windle (2013). "Bayesian inference for logistic models using Pólya-Gamma latent variables". In: *Journal of the American Statistical Association* 108.504, pp. 1339–1349. ISSN: 1537274X. DOI: 10.1080/01621459.2013.829001. arXiv: 1205.0310.

Puranam, Dinesh, Vishal Narayan, and Vrinda Kadiyali (2017). "The effect of calorie posting regulation on consumer opinion: A flexible latent dirichlet allocation model with informative priors". In: *Marketing Science* 36.5, pp. 726–746. ISSN: 1526548X. DOI: 10.1287/mksc.2017.1048.

Qi, Jiayin et al. (2016). "Mining customer requirements from online reviews: A product improvement perspective". In: *Information and Management* 53.8, pp. 951–963. ISSN: 03787206. DOI: 10.1016/j.im.2016.06.002. URL: http://dx.doi.org/10.1016/j.im.2016.06.002.

Ramage, Daniel et al. (2009). "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora". In: *Empirical Methods in Natural Language Processing*. August, pp. 248–256. URL: http://www.aclweb.org/anthology/

D09-1026{\%}5Cnpapers2://publication/uuid/1C96B11C-B699-460E-8674-E4B40B34AE20.

Rand, William and Roland T. Rust (2011). "Agent-based modeling in marketing: Guidelines for rigor". In: *International Journal of Research in Marketing* 28.3, pp. 181–193. ISSN: 01678116. DOI: 10.1016/j.ijresmar.2011.04.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167811611000504.

Risselada, Hans, Peter C. Verhoef, and Tammo H.A. Bijmolt (2016). "Indicators of opinion leadership in customer networks: self-reports and degree centrality". In: *Marketing Letters* 27.3, pp. 449–460. ISSN: 09230645. DOI: 10.1007/s11002-015-9369-7. URL: http://dx.doi.org/10.1007/s11002-015-9369-7.

Rossi, Peter E., Greg M. Allenby, and Robert McCulloch (2005). *Bayesian Statistics and Marketing*. Vol. 1. Wiley Series in Probability and Statistics 3. Chichester, UK: John Wiley & Sons, Ltd. ISBN: 9780470863695. DOI: 10.1002/0470863692. URL: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470863692{\%}0Ahttp://search.ebscohost.com/login.aspx?direct=true{\&}db=bth{\&}AN=6025342{\&}site=ehost-livehttp://doi.wiley.com/10.1002/0470863692.

Saldaña, D. Franco, Yi Yu, and Yang Feng (2017). "How Many Communities Are There?" In: *Journal of Computational and Graphical Statistics* 26.1, pp. 171–181. ISSN: 1061-8600. DOI: 10.1080/10618600.2015.1096790. URL: https://www.tandfonline.com/doi/full/10.1080/10618600.2015.1096790.

Schulze, Christian, Lisa Schöler, and Bernd Skiera (2014). "Not all fun and games: Viral marketing for utilitarian products". In: *Journal of Marketing* 78.1, pp. 1–19. ISSN: 15477185. DOI: 10.1509/jm.11.0528.

Snijders, Tom A.B. and Krzysztof Nowicki (1997). "Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure". In: *Journal of Classification* 14.1, pp. 75–100. ISSN: 0176-4268. DOI: 10.1007/s003579900004. URL: http://link.springer.com/10.1007/s003579900004.

Sonnier, Garrett P., Leigh Mcalister, and Oliver J. Rutz (2011). "A dynamic model of the effect of online communications on firm sales". In: *Marketing Science* 30.4, pp. 702–716. ISSN: 07322399. DOI: 10.1287/mksc.1110.0642.

Spiegelhalter, David J. et al. (2002). "Bayesian measures of model complexity and fit". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.4, pp. 583–639. ISSN: 1369-7412. DOI: 10.1111/1467-9868.00353. URL: http://doi.wiley.com/10.1111/1467-9868.00353.

Stekhoven, D. J. and P. Buhlmann (2012). "MissForest–non-parametric missing value imputation for mixed-type data". In: *Bioinformatics* 28.1, pp. 112–118. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr597. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr597.

Susarla, Anjana, Jeong Ha Oh, and Yong Tan (2012). "Social networks and the diffusion of user-generated content: Evidence from youtube". In: *Information Systems Research* 23.1, pp. 23–41. ISSN: 15265536. DOI: 10.1287/isre.1100.0339.

Tirunillai, Seshadri and Gerard J. Tellis (2014). "Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation". In: *Journal of Marketing Research* 51.4, pp. 463–479. ISSN: 00222437. DOI: 10.1509/jmr.12.0106.

Toker-Yildiz, Kamer et al. (2017). "Social Interactions and Monetary Incentives in Driving Consumer Repeat Behavior". In: *Journal of Marketing Research* 54.3, pp. 364–380. ISSN: 0022-2437. DOI: 10.1509/jmr.13.0482. URL: http://journals.sagepub.com/doi/10.1509/jmr.13.0482.

Toubia, Olivier et al. (2019). "Extracting Features of Entertainment Products: A Guided Latent Dirichlet Allocation Approach Informed by the Psychology of Media Consumption". In: *Journal of Marketing Research* 56.1, pp. 18–36. DOI: 10.1177/0022243718820559. URL: http://journals.sagepub.com/doi/10.1177/0022243718820559.

Trusov, Michael, Anand V. Bodapati, and Randolph E. Bucklin (2010). "Determining Influential Users in Internet Social Networks". In: *Journal of Marketing Research* 47.4, pp. 643–658. ISSN: 0022-2437. DOI: 10.1509/jmkr.47.4.643. URL: http://journals.sagepub.com/doi/10.1509/jmkr.47.4.643.

Wang, Jing, Anocha Aribarg, and Yves F. Atchadé (2013). "Modeling choice interdependence in a social network". In: *Marketing Science* 32.6, pp. 977–997. ISSN: 1526548X. DOI: 10.1287/mksc.2013.0811.

Wang, Yuchung J. and George Y. Wong (1987). "Stochastic Blockmodels for Directed Graphs". In: *Journal of the American Statistical Association* 82.397, pp. 8–19.

Watanabe, Sumio (2010). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory". In: *Journal of Machine Learning Research* 11, pp. 3571–3594. ISSN: 15324435. arXiv: 1004.2316.

Wu, Chunhua et al. (2015). "The economic value of online reviews". In: *Marketing Science* 34.5, pp. 739–754. ISSN: 1526548X. DOI: 10.1287/mksc.2015.0926.

Xiao, Shengsheng, Chih Ping Wei, and Ming Dong (2016). "Crowd intelligence: Analyzing online product reviews for preference measurement". In: *Information and Management* 53.2, pp. 169–182. ISSN: 03787206. DOI: 10.1016/j.im.2015.09.010. URL: http://dx.doi.org/10.1016/j.im.2015.09.010.

Xing, Eric P., Wenjie Fu, and Le Song (2010). "A state-space mixed membership blockmodel for dynamic network tomography". In: *The Annals of Applied Statistics* 4.2, pp. 535–566. ISSN: 1932-6157. DOI: 10.1214/09-AOAS311. URL: http://projecteuclid.org/euclid.aoas/1280842130.

Xu, Kevin S. and Alfred O. Hero (2014). "Dynamic Stochastic Blockmodels for Time-Evolving Social Networks". In: *IEEE Journal of Selected Topics in Signal Processing* 8.4, pp. 552–562. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2014.2310294. URL: https://ieeexplore.ieee.org/document/6758385/.

Yin, Dezhi, Samuel D. Bond, and Han Zhang (2017). "Keep Your Cool or Let it Out: Nonlinear Effects of Expressed Arousal on Perceptions of Consumer Reviews". In: *Journal of Marketing Research* 54.3, pp. 447–463. ISSN: 0022-2437. DOI: 10.1509/jmr.13.0379. URL: http://journals.sagepub.com/doi/10.1509/jmr.13.0379.

Yoganarasimhan, Hema (2012). *Impact of social network structure on content propagation: A study using YouTube data*. Vol. 10. 1, pp. 111–150. ISBN: 1112901191. DOI: 10.1007/s11129-011-9105-4.

Zanghi, Hugo, Stevenn Volant, and Christophe Ambroise (2010). "Clustering based on random graph model embedding vertex features". In: *Pattern Recognition Letters* 31.9, pp. 830–836. ISSN: 01678655. DOI: 10.1016/j.patrec.2010.01.026. URL: http://dx.doi.org/10.1016/j.patrec.2010.01.026.

Zhang, Dongwen et al. (2015). "Chinese comments sentiment classification based on word2vec and SVMperf". In: *Expert Systems with Applications* 42.4, pp. 1857–1863. ISSN: 09574174. DOI: 10.1016/j.eswa.2014.09.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417414005508.

Zhang, Honghong et al. (2018). “When are influentials equally influenceable? The strength of strong ties in new product adoption”. In: *Journal of Business Research* 82.September 2017, pp. 160–170. ISSN: 01482963. DOI: 10.1016/j.jbusres.2017.09.013.

Zhang, Jonathan Z. (2019). “Dynamic customer interdependence”. In: *Journal of the Academy of Marketing Science* 47.4, pp. 723–746. ISSN: 0092-0703. DOI: 10.1007/s11747-019-00627-z. URL: http://link.springer.com/10.1007/s11747-019-00627-z.

Zhang, Yuchi and David Godes (2018). “Learning from online social ties”. In: *Marketing Science* 37.3, pp. 425–444. ISSN: 1526548X. DOI: 10.1287/mksc.2017.1076.

Zhang, Yuchi, Wendy W. Moe, and David A. Schweidel (2017). “Modeling the role of message content and influencers in social media rebroadcasting”. In: *International Journal of Research in Marketing* 34.1, pp. 100–119. ISSN: 01678116. DOI: 10.1016/j.ijresmar.2016.07.003. URL: http://dx.doi.org/10.1016/j.ijresmar.2016.07.003.

Zhu, Feng and Xiaoquan Zhang (2010). “Impact of online consumer reviews on Sales: The moderating role of product and consumer characteristics”. In: *Journal of Marketing* 74.2, pp. 133–148. ISSN: 00222429. DOI: 10.1509/jmkg.74.2.133.

Zhu, Yaojia et al. (2013). “Scalable text and link analysis with mixed-Topic link models”. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 473–481. DOI: 10.1145/2487575.2487693. arXiv: 1303.7264.